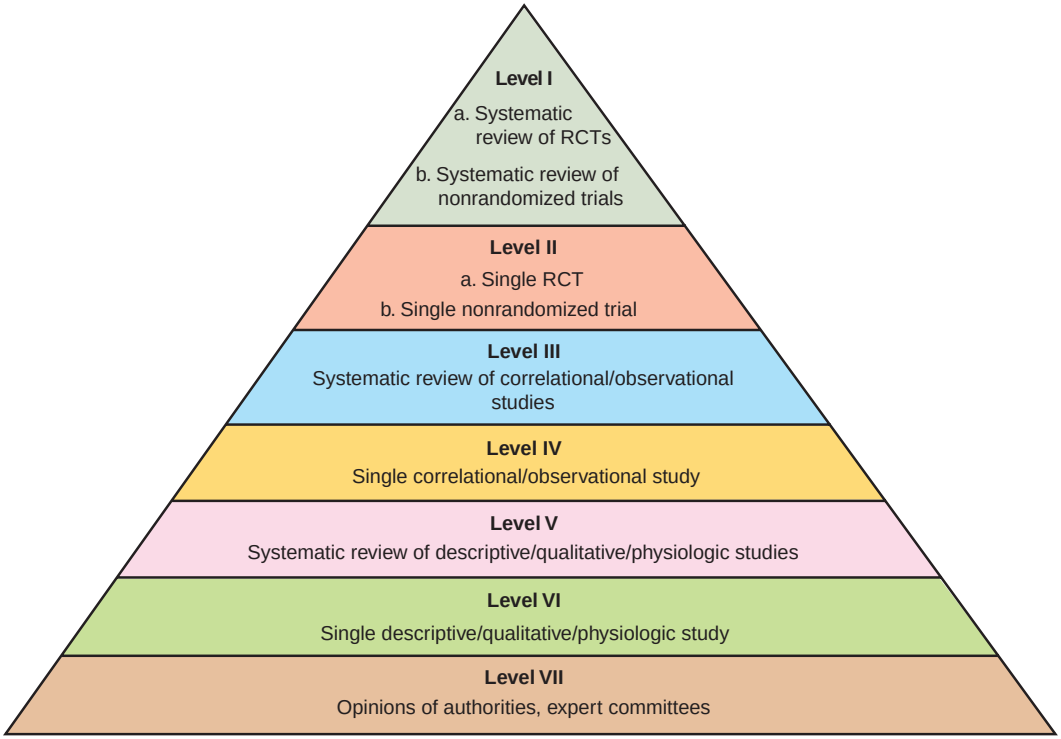
QUANTITATIVE HEALTH RESEARCH

Quick Guide to an Evidence Hierarchy of Designs for Cause-Probing Questions



Contents

Quick Guide to Bivariate Statistical Tests

Level of measurement of dependent variable	Group Comparisons: Number of groups (the independent variable)				Correlational analyses (To examine relationship strength)
	2 Groups		3+ Groups		
	Independent Groups Tests	Dependent Groups Tests	Independent Groups Tests	Dependent Groups Tests	
Nominal (Categorical)	χ^2 pp. 420–421 (or Fisher's exact test) p. 421	McNemar"s test p. 421	χ^2 pp. 420–421	Cochran's *Q*	Phi coefficient (dichotomous) or Cramér's *V* (not restricted to dichotomous) p. 422
Ordinal (Rank)	Mann-Whitney Test (or Median test) p. 416	Wilcoxon signed ranks test p. 416	Kruskal-Wallis *H* test p. 420	Friedman's test p. 420	Spearman's rho (or Kendall's tau) p. 422
Interval or Ratio (Continuous)*	Independent group *t test* pp. 413–415	Paired *t test* p. 415	ANOVA pp. 416–420	RM-ANOVA p. 420	Pearson's *r* pp. 421–422
	Multifactor ANOVA for 2+ independent variables p. 420				
	RM-ANOVA for 2+ groups x 2+ measurements over time p. 446				

*For distributions that are markedly nonnormal or samples that are small, the nonparametric tests in the row above may be needed.

3 Key Concepts and Steps in Qualitative and Quantitative Research

This chapter covers a lot of ground—but, for many of you, it is familiar ground. For those who have taken an earlier research course, this chapter provides a review of key terms and steps in the research process. For those without previous exposure to research methods, this is an important chapter that offers basic grounding in research terminology.

Research, like any discipline, has its own language—its own *jargon.* Some terms are used by both qualitative and quantitative researchers, but others are used predominantly by one or the other group. To make matters more complex, much of the jargon used in nursing research has its roots in the social sciences, but sometimes different terms for the same concepts are used in medical research; we cover both but acknowledge that social science jargon predominates.

FUNDAMENTAL RESEARCH TERMS AND CONCEPTS

When researchers address a problem through research—regardless of the underlying paradigm—they undertake a **study** (or an **investigation**). Studies involve various people working together in different roles.

The Faces and Places of Research

Studies with humans involve two sets of people: those who do the research and those who provide the information. In a quantitative study, the people being studied are called **subjects** or **study participants (**Table 3.1). In a qualitative study, the individuals cooperating in the study are called **informants**, **key informants**, or study participants. Collectively, both in qualitative and quantitative studies, study participants comprise the **sample**.

The person who conducts a study is the **researcher** or **investigator**. Studies are often undertaken by several people. When a study is done by a team, the person directing the study is the **principal investigator (PI)**. Two or three researchers collaborating equally are **co-investigators**. **Reviewers** are sometimes called on to critique a study and offer feedback. If these people are at a similar level of experience to the researchers, they are **peer reviewers**.

In large-scale projects, dozens of individuals may be involved in planning, managing, and conducting the study. The examples of staffing configurations that follow span the continuum from an extremely large project to a more modest one.

Examples of staffing on a quantitative study: The first author of this book was involved in a multicomponent, interdisciplinary study of poor

TABLE 3.1 Key Terms in Quantitative and Qualitative Research

CONCEPT	QUANTITATIVE TERM	QUALITATIVE TERM
Person Contributing Information	Subject Study participant —	— Study participant Informant, key informant
Person Undertaking the Study	Researcher Investigator	Researcher Investigator
That Which Is Being Investigated	— Concepts Constructs Variables	Phenomena Concepts — —
System of Organizing Concepts	Theory, theoretical framework Conceptual framework, conceptual model	Theory Conceptual framework, sensitizing framework
Information Gathered	Data (numerical values)	Data (narrative descriptions)
Connections Between Concepts	Relationships (cause-and-effect, functional)	Patterns of association
Logical reasoning processes	Deductive reasoning	Inductive reasoning

women living in four major cities (Cleveland, Los Angeles, Miami, and Philadelphia). As part of the study, she and two colleagues prepared a report documenting the health problems of about 4,000 welfare mothers who were interviewed in 1998 and again in 2001 (Polit et al., 2001). The project staff included over 100 people, including 2 co-PIs; lead investigators (Polit was one) of 6 project components; over 50 interviewers and supervisors; and dozens of other researchers, research assistants, computer programmers, and other support staff. Several health consultants, including a prominent nurse researcher (Linda Aiken), were reviewers.

Examples of staffing on a qualitative study: Beck (2009) conducted a qualitative study focusing on the experiences of mothers caring for their children with a brachial plexus injury. The team consisted of Beck as the PI (who gathered and analyzed all the data), members of the United Brachial Plexus Executive Board (who helped to recruit mothers for the study), a transcriber (who listened to the tape-recorded interviews and typed them up verbatim), and an undergraduate nursing student (who checked the accuracy of the interview transcripts against the tape-recorded interviews). (Beck's study appears in its entirety in the accompanying *Resource Manual*).

Research can be undertaken in a variety of *settings* (the specific places where information is gathered), and in one or more *sites*. Some studies take place in **naturalistic settings** in the field, such as in people's homes, but some studies are done in controlled **laboratory settings**. Researchers make decisions about where to conduct a study based on the nature of the research question and type of information needed. Qualitative researchers are especially likely to engage in **fieldwork** in natural

settings because they are interested in the contexts of people's experiences. The *site* is the overall location for the research—it could be an entire community (e.g., a Haitian neighborhood in Miami) or an institution (e.g., a hospital in Toronto). Researchers sometimes engage in **multisite studies** because the use of multiple sites offers a larger or more diverse sample of study participants.

The Building Blocks of Research

Phenomena, Concepts, and Constructs

Research involves abstractions. For example, *pain*, *quality of life*, and *resilience* are abstractions of particular aspects of human behavior and characteristics. These abstractions are called **concepts** or, in qualitative studies, **phenomena**.

Researchers may also use the term **construct**. Like a concept, a construct is an abstraction inferred from situations or behaviors. Kerlinger and Lee (2000) distinguish concepts from constructs by noting that constructs are abstractions that are deliberately and systematically invented (constructed) by researchers. For example, *self-care* in Orem's model of health maintenance is a construct. The terms *construct* and *concept* are sometimes used interchangeably but, by convention, a construct refers to a more complex abstraction than a concept.

Theories and Conceptual Models

A **theory** is a systematic, abstract explanation of some aspect of reality. Theories, which knit concepts together into a coherent system, play a role in both qualitative and quantitative research.

Quantitative researchers may start with a theory, *framework*, or *conceptual model* (distinctions are discussed in Chapter 6). Based on theory, they make predictions about how phenomena will behave in the real world *if the theory is true*. Specific predictions deduced from theory are tested through research; results are used to support, reject, or modify the theory.

In qualitative research, theories may be used in various ways. Sometimes conceptual or **sensitizing frameworks**, derived from qualitative research traditions we describe later in this chapter, provide an impetus for a study or offer an orienting world view. In such studies, the framework helps to guide the inquiry and to interpret gathered information. In other qualitative studies, theory is the *product* of the research: The investigators use information from participants inductively to develop a theory rooted in the participants' experiences. The goal is to develop a theory that explains phenomena *as they exist*, not as they are preconceived.

Variables

In quantitative studies, concepts are usually called **variables**. A variable, as the name implies, is something that varies. Weight, anxiety, and blood pressure are variables—each varies from one person to another. In fact, most aspects of humans are variables. If everyone weighed 150 pounds, weight would not be a variable, it would be a *constant*. It is precisely because people and conditions *do* vary that most research is conducted. Quantitative researchers seek to understand how or why things vary, and to learn if differences in one variable are related to differences in another. For example, lung cancer research is concerned with the variable of lung cancer, which is a variable because not everyone has this disease. Researchers have studied factors that might be linked to lung cancer, such as cigarette smoking. Smoking is also a variable because not everyone smokes. A variable, then, is any quality of a person, group, or situation that varies or takes on different values. Variables are the building blocks of quantitative studies.

When an attribute is extremely varied in the group under study, the group is **heterogeneous** with respect to that variable. If the amount of variability is limited, the group is **homogeneous**. For example, for the variable height, a group of 2-year-old children is likely to be more homogeneous than a group of 18-year-olds. Degree of **variability** or **heterogeneity** of a group of people has implications for study design.

Variables may be inherent characteristics of people, such as their age, blood type, or weight. Sometimes, however, researchers *create* a variable. For example, if a researcher tests the effectiveness of patient-controlled analgesia as opposed to

intramuscular analgesia in relieving pain after surgery, some patients would be given patient-controlled analgesia and others would receive intramuscular analgesia. In the context of this study, method of pain management is a variable because different patients get different analgesic methods.

Continuous, Discrete, and Categorical Variables. Some variables take on a wide range of values. A person's age, for instance, can take on values from zero to more than 100, and the values are not restricted to whole numbers. **Continuous variables** have values along a continuum and, in theory, can assume an infinite number of values between two points. Consider the continuous variable *weight*: between 1 and 2 pounds, the number of values is limitless: 1.05, 1.8, 1.333, and so on.

By contrast, a **discrete variable** has a finite number of values between any two points, representing discrete quantities. For example, if people were asked how many children they had, they might answer 0, 1, 2, 3, or more. The value for number of children is discrete, because a number such as 1.5 is not meaningful. Between 1 and 3, the only possible value is 2.

Other variables take on a small range of values that do not represent a *quantity*. Blood type, for example, has four values—A, B, AB, and O. Variables that take on a handful of discrete nonquantitative values are **categorical variables**. When categorical variables take on only two values, they are **dichotomous variables**. Gender, for example, is dichotomous: male and female.

Dependent and Independent Variables. Many studies seek to unravel and understand causes of phenomena. Does a nursing intervention *cause* improvements in patient outcomes? Does smoking *cause* lung cancer? The presumed cause is the **independent variable**, and the presumed effect is the **dependent variable**. Some researchers use the term **outcome variable**—the variable capturing the outcome of interest—in lieu of dependent variable.

Variability in the dependent variable is presumed to *depend on* variability in the independent variable. For example, researchers study the extent to which lung cancer (the dependent variable) depends on smoking (the independent variable). Or, investigators may study the extent to which patients' pain (the dependent variable) depends on different nursing actions (the independent variable).

Frequently, the terms *independent variable* and *dependent variable* are used to indicate *direction of influence* rather than a causal mechanism. For example, suppose a researcher studied the mental health of caregivers caring for spouses with Alzheimer's disease and found better mental health outcomes for wives than for husbands. The researcher might be unwilling to conclude that caregivers' mental health was *caused* by gender. Yet the direction of influence clearly runs from gender to mental health: It makes *no* sense to suggest that caregivers' mental health influenced their gender! Although the researcher cannot infer a cause-and-effect connection, it is appropriate to conceptualize mental health as the dependent variable and gender as the independent variable, because it is the caregivers' mental health that the researcher is interested in understanding, explaining, or predicting.

Most dependent variables have multiple causes or antecedents. If we were studying factors that influence people's weight, we might consider their height, physical activity, and diet as independent variables. Two or more *dependent* variables also may be of interest. For example, a researcher may compare the effects of two methods of nursing care for children with cystic fibrosis. Several dependent variables could be used to assess treatment effectiveness, such as length of hospital stay, number of recurrent respiratory infections, and so on. It is common to design studies with multiple independent and dependent variables.

Variables are not *inherently* dependent or independent. A dependent variable in one study could be an independent variable in another. For example, a study might examine the effect of a nurse-initiated exercise intervention (the independent variable) on osteoporosis (the dependent variable). Another study might investigate the effect of osteoporosis (the independent variable) on bone fracture incidence (the dependent variable). In short, whether a

variable is independent or dependent is a function of the role that it plays in a particular study.

Example of independent and dependent variables: *Research question:* Do women with diabetes differ from those without diabetes in terms of cancer screening behaviors? (Marshall et al., 2010)
Independent variable: Status of having or not having diabetes
Dependent variable: Cancer screening behaviors

Conceptual and Operational Definitions

Study concepts need to be defined and explicated, and dictionary definitions are seldom adequate. Two types of definitions are of particular relevance—conceptual and operational.

Concepts are abstractions of observable phenomena, and researchers' world views shapes how those concepts are defined. A **conceptual definition** presents the abstract or theoretical meaning of the concepts being studied. Even seemingly straightforward terms need to be conceptually defined. The classic example is the concept of *caring*. Morse and colleagues (1990) scrutinized the works of numerous writers to determine how *caring* was defined, and identified five different classes of conceptual definition: as a human trait, a moral imperative, an affect, an interpersonal relationship, and a therapeutic intervention. Researchers undertaking studies concerned with caring need to make clear which conceptual definition they have adopted—both to themselves and to their readers. In qualitative studies, conceptual definitions of key phenomena may be the major end product of the endeavor, reflecting the intent to have the meaning of concepts defined by those being studied.

In quantitative studies, however, researchers clarify and define concepts at the outset. This is necessary because quantitative researchers must indicate how the variables will be observed and measured. An **operational definition** of a concept specifies the operations that researchers must perform to measure it. Operational definitions should be congruent with conceptual definitions.

Variables differ in the ease with which they can be operationalized. The variable *weight*, for example, is easy to define and measure. We might operationally define weight as the amount that an object weighs, to the nearest full pound. This definition designates that weight will be measured using one system (pounds) rather than another (grams). We could also specify that weight will be measured using a spring scale with participants fully undressed after 10 hours of fasting. This operational definition clearly indicates what we mean by the variable *weight*.

Few variables are operationalized as easily as weight. Most variables can be measured in different ways, and researchers must choose the one that best captures the variables as they conceptualize them. Take, for example, *anxiety*, which can be defined in terms of both physiologic and psychological functioning. For researchers choosing to emphasize physiologic aspects, the operational definition might involve a physiologic measure such as the Palmar Sweat Index. If researchers conceptualize anxiety as a psychological state, the operational definition might involve a paper-and-pencil measure such as the State Anxiety Scale. Readers of research articles may not agree with how variables were conceptualized and measured, but definitional precision has the advantage of communicating exactly what terms mean within the study.

Example of conceptual and operational definitions: Schim, Doorenbos, and Borse (2006) tested an intervention to expand cultural competence among hospice workers. Cultural competence encompassed several aspects, such as cultural awareness, which was conceptually defined as a care provider's knowledge about areas of cultural expression in which cultural groups may differ. The researchers measured their constructs with the Cultural Competence Assessment (CCA) instrument. The CCA operationalizes cultural awareness by having healthcare staff indicate their level of agreement with such statements as, "I understand that people from different cultural groups may define the concept of 'healthcare' in different ways."

Data

Research **data** (singular, datum) are the pieces of information obtained in a study. In quantitative

BOX 3.1 Example of Quantitative Data

Question: Thinking about the past week, how depressed would you say you have been on a scale from 0 to 10, where 0 means "not at all" and 10 means "the most possible"?

Data: 9 (Subject 1)
0 (Subject 2)
4 (Subject 3)

studies, researchers identify variables, develop conceptual and operational definitions, and then collect relevant data. Quantitative researchers collect primarily **quantitative data**—data in numeric form. For example, suppose we conducted a quantitative study in which a key variable was depression. We might ask, "Thinking about the past week, how depressed would you say you have been on a scale from 0 to 10, where 0 means 'not at all' and 10 means 'the most possible'?" Box 3.1 presents quantitative data for three fictitious people. Subjects provided a number along the 0 to 10 continuum representing their degree of depression—*9* for subject 1 (a high level of depression), *0* for subject 2 (no depression), and *4* for subject 3 (little depression). The numeric values for all people, collectively, would comprise the data on depression.

In qualitative studies, researchers collect **qualitative data**, that is, narrative descriptions. Narrative information can be obtained by having conversations with participants, by making detailed notes about how people behave in naturalistic settings, or by obtaining narrative records, such as diaries. Suppose we were studying depression qualitatively. Box 3.2 presents qualitative data for three people responding conversationally to the question, "Tell me about how you've been feeling lately—have you felt sad or depressed at all, or have you generally been in good spirits?" The data consist of rich descriptions of each participant's emotional state.

Relationships

Researchers are rarely interested in isolated concepts, except in descriptive studies. For example, a researcher might describe the percentage of patients receiving intravenous (IV) therapy who experience IV infiltration. In this example, the variable is IV

BOX 3.2 Example of Qualitative Data

Question: Tell me about how you've been feeling lately—have you felt sad or depressed at all, or have you generally been in good spirits?

Data: "Well, actually, I've been pretty depressed lately, to tell you the truth. I wake up each morning and I can't seem to think of anything to look forward to. I mope around the house all day, kind of in despair. I just can't seem to shake the blues, and I've begun to think I need to go see a shrink." (Participant 1)

"I can't remember ever feeling better in my life. I just got promoted to a new job that makes me feel like I can really get ahead in my company. And I've just gotten engaged to a really great guy who is very special." (Participant 2)

"I've had a few ups and downs the past week, but basically things are on a pretty even keel. I don't have too many complaints." (Participant 3)

infiltration versus no infiltration. Usually, however, researchers study phenomena in relation to other phenomena—that is, they focus on relationships. A **relationship** is a bond or a connection between phenomena. For example, researchers repeatedly have found a *relationship* between cigarette smoking and lung cancer. Both qualitative and quantitative studies examine relationships, but in different ways.

In quantitative studies, researchers examine the relationship between the independent and dependent variables. The research question asks whether variation in the dependent variable is systematically related to variation in the independent variable. Relationships are usually expressed in quantitative terms, such as *more than*, *less than*, and so on. For example, let us consider as our dependent variable a person's weight. What variables are related to (associated with) body weight? Some possibilities are height, caloric intake, and exercise. For each independent variable, we can make a prediction about its relationship to the dependent variable:

Height: Taller people will weigh more than shorter people.
Caloric intake: People with higher caloric intake will be heavier than those with lower caloric intake.
Exercise: The lower the amount of exercise, the greater will be the person's weight.

Each statement expresses a predicted relationship between weight (the dependent variable) and a measurable independent variable. Terms such as *more than* and *heavier than* imply that as we observe a change in one variable, we are likely to observe a change in weight. If Nate were taller than Tom, we would predict (in the absence of any other information) that Nate is also heavier than Tom.

Quantitative studies can address one or more of the following questions about relationships:

- Does a relationship between variables *exist*? (e.g., is cigarette smoking related to lung cancer?)
- What is the *direction* of the relationship between variables? (e.g., are people who smoke *more* likely or *less* likely to get lung cancer than those who do not?)
- How *strong* is the relationship between the variables? (e.g., how powerful is the link between smoking and lung cancer? How much higher is the risk that smokers will develop lung cancer?)
- What is the *nature* of the relationship between variables? (e.g., does smoking *cause* lung cancer? Does some other factor *cause* both smoking and lung cancer?)

As the last question suggests, variables can be related to one another in different ways. One type of relationship is called a **cause-and-effect** (or **causal**) **relationship**. Within the positivist paradigm, natural phenomena are assumed not to be haphazard; they have antecedent causes that are presumably discoverable. In our example about a person's weight, we might speculate that there is a causal relationship between caloric intake and weight: consuming more calories causes weight gain. As noted in Chapter 1, many quantitative studies are *cause-probing*—they seek to illuminate the causes of phenomena.

Example of a study of causal relationships: Lin and colleagues (2010) studied whether a therapeutic lifestyle program caused reductions in cardiac risk factors following coronary artery bypass graft surgery.

Not all relationships between variables can be interpreted as cause-and-effect relationships. There is a relationship, for example, between a person's pulmonary artery and tympanic temperatures: people with high readings on one tend to have high readings on the other. We cannot say, however, that pulmonary artery temperature *caused* tympanic temperature, nor that tympanic temperature *caused* pulmonary artery temperature. This type of relationship is called a **functional** (or an **associative**) **relationship** rather than a causal relationship.

Example of a study of functional relationships: Al-Akour and co-researchers (2010) examined the relationship between quality of life among Jordanian adolescents with type 1 diabetes on the one hand, and gender and age on the other.

Qualitative researchers are not concerned with quantifying relationships, nor in testing causal

relationships. Qualitative researchers seek patterns of association as a way to illuminate the underlying meaning and dimensionality of phenomena. Patterns of interconnected themes and processes are identified as a means of understanding the whole.

Example of a qualitative study of patterns: Gaudine and colleagues (2010) studied HIV-related stigma in a Vietnamese community. In-depth interviews were conducted with people living with HIV, family members, community members, and healthcare professionals. The researchers identified four dimensions of HIV-related stigma, the manifestation of which differed for each group.

MAJOR CLASSES OF QUANTITATIVE AND QUALITATIVE RESEARCH

Researchers usually work within a paradigm that is consistent with their world view, and that gives rise to questions that excite their curiosity. The maturity of the focal concept also may lead to one or the other paradigm: When little is known about a topic, a qualitative approach is often more fruitful than a quantitative one. In this section, we briefly describe broad categories of quantitative and qualitative research.

Quantitative Research: Experimental and Nonexperimental Studies

A basic distinction in quantitative studies is between experimental and nonexperimental research. In **experimental research**, researchers actively introduce an intervention or treatment. In **nonexperimental research**, researchers are bystanders—they collect data without intervening. For example, if a researcher gave bran flakes to one group of people and prune juice to another to evaluate which method facilitated elimination more effectively, the study would be experimental because the researcher intervened in the normal course of things. If, however, a researcher compared elimination patterns of two groups of people whose regular eating patterns differed—for example, some normally took foods that stimulated bowel elimination and others did not—there is no intervention, and the study is nonexperimental. In medical and epidemiologic research, an experimental study usually is called a **clinical trial**, and a nonexperimental inquiry is called an **observational study**. As we discuss in Chapter 11, a *randomized controlled trial* or RCT is a particular type of clinical trial.

Experimental studies are explicitly cause-probing—they test whether an intervention *caused* changes in (affected) the dependent variable. Sometimes nonexperimental studies also seek to elucidate or detect causal relationships, but the resulting evidence is usually less conclusive. Experimental studies offer the possibility of greater control over confounding influences than nonexperimental studies, and so, causal inferences are more plausible.

Example of experimental research: Twiss and colleagues (2009) tested the effect of an exercise intervention for breast cancer survivors with bone loss on the women's muscle strength, balance, and fall frequency. Some women received the 24-month intervention, and others did not.

In this example, the researcher intervened by giving some patients the opportunity to participate in the exercise program, while others were not given this opportunity. In other words, the researcher *controlled* the independent variable, which in this case was the exercise intervention.

Example of nonexperimental research: Vallance and co-researchers (2010) studied factors that predicted exercise and physical activity among breast cancer survivors. They examined the association between physical activity on the one hand and demographic, psychosocial, and motivational factors measured 6 months earlier on the other.

This nonexperimental study did not involve an intervention. The researchers were interested in similar variables as in the previously described experimental study (physical activity and exercise) and in a similar population (patients with breast cancer), but their intent was to explore existing relationships rather than to evaluate an intervention.

Qualitative Research: Disciplinary Traditions

The majority of qualitative studies can best be described as **qualitative descriptive research**. Many qualitative studies, however, are rooted in research traditions that originated in anthropology, sociology, and psychology. Three such traditions, prominent in qualitative nursing research, are briefly described here. Chapter 19 provides a fuller discussion of these traditions and the methods associated with them.

The **grounded theory** tradition, with roots in sociology, seeks to describe and understand the key social psychological processes that occur in a social setting. Grounded theory was developed in the 1960s by two sociologists, Glaser and Strauss (1967). The focus of most grounded theory studies is on a developing social experience—the social and psychological stages and phases that characterize a particular event or episode. A major component of grounded theory is the discovery of a *core variable* that is central in explaining what is going on in that social scene. Grounded theory researchers strive to generate explanations of phenomena that are grounded in reality.

Example of a grounded theory study: Propp and colleagues (2010) conducted a grounded theory study to examine critical healthcare team processes. They identified specific nurse–team communication practices that were perceived by team members to enhance patient outcomes.

Phenomenology, rooted in a philosophical tradition developed by Husserl and Heidegger, is concerned with the lived experiences of humans. Phenomenology is an approach to thinking about what life experiences of people are like and what they mean. The phenomenological researcher asks the questions: What is the *essence* of this phenomenon as experienced by these people? Or, what is the meaning of the phenomenon to those who experience it?

Example of a phenomenological study: Schachman (2010) conducted in-depth interviews to explore the lived experience of first-time fatherhood from the perspective of military men deployed to combat regions during birth.

Ethnography is the primary research tradition within anthropology, and provides a framework for studying the lifeways and experiences of a defined cultural group. Ethnographers typically engage in extensive fieldwork, often participating in the life of the culture under study. Ethnographic research is in some cases concerned with broadly defined cultures (e.g., Hmong refugee communities), but sometimes focuses on more narrowly defined cultures (e.g., the culture of an emergency department). Ethnographers strive to learn from members of a cultural group, to understand their world view, and to describe their customs and norms.

Example of an ethnographic study: Hessler (2009) conducted ethnographic fieldwork to investigate physical activity and active play among rural preschool children.

MAJOR STEPS IN A QUANTITATIVE STUDY

In quantitative studies, researchers move from the beginning of a study (posing a question) to the end point (obtaining an answer) in a reasonably linear sequence of steps that are broadly similar across studies. In some studies, the steps overlap; in others, certain steps are unnecessary. Still, a general flow of activities is typical in a quantitative study (See Figure 3.1). This section describes that flow, and the next section describes how qualitative studies differ.

Phase 1: The Conceptual Phase

Early steps in a quantitative study typically have a strong conceptual or intellectual element. These activities include reading, conceptualizing, theorizing, and reviewing ideas with colleagues or advisers. During this phase, researchers call on such skills as creativity, deductive reasoning, and a firm grounding in previous research on the topic of interest.

Step 1: Formulating and Delimiting the Problem

Quantitative researchers begin by identifying an interesting, significant research problem and

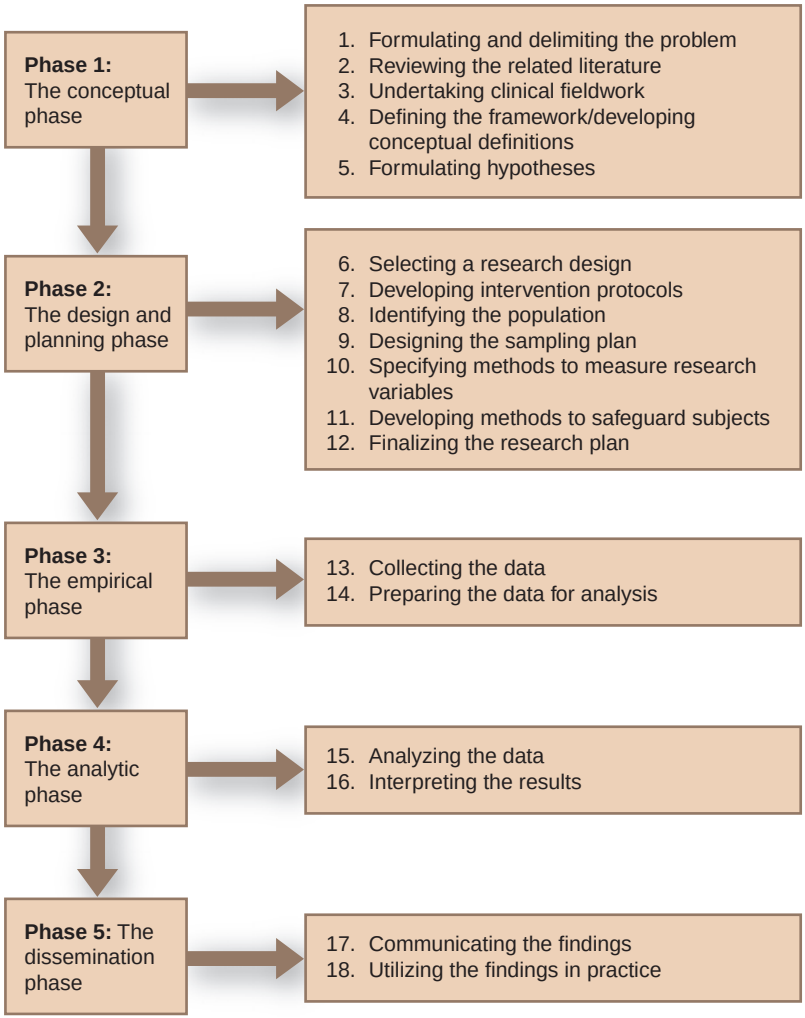


FIGURE 3.1 Flow of steps in a quantitative study.

formulating **research questions**. Good research depends to a great degree on good questions. In developing research questions, nurse researchers must attend to substantive issues (What kind of new evidence is needed?), theoretical issues (Is there a conceptual context for understanding this problem?), clinical issues (How could evidence from this study be used in clinical practice?), methodologic issues (How can this question best be studied to yield high-quality evidence?), and ethical issues (Can this question be rigorously addressed without committing ethical transgressions?).

TIP: A critical ingredient in developing good research questions is personal interest. Begin with topics that fascinate you or about which you have a passionate interest or curiosity.

Step 2: Reviewing the Related Literature

Quantitative research is typically conducted in the context of previous knowledge. To contribute new

evidence, quantitative researchers strive to understand existing evidence. A thorough **literature review** provides a foundation on which to base new evidence and usually is conducted before data are collected. For clinical problems, it may also be necessary to learn the "status quo" of current procedures, and to review existing practice guidelines or protocols.

Step 3: Undertaking Clinical Fieldwork

Unless the research problem originated in a clinical setting, researchers embarking on a clinical nursing study benefit from spending time in clinical settings, discussing the problem with clinicians and administrators, and observing current practices. Clinical fieldwork can provide perspectives on recent clinical trends, current diagnostic procedures, and relevant healthcare-delivery models; it can also help researchers better understand clients and the settings in which care is provided. Such fieldwork can also be valuable in gaining access to an appropriate site or in developing methodologic strategies. For example, in the course of clinical fieldwork researchers might discover the need for research assistants who are bilingual.

Step 4: Defining the Framework and Developing Conceptual Definitions

Theory is the ultimate aim of science: It transcends the specifics of a particular time, place, and group and aims to identify regularities in the relationships among variables. When quantitative research is performed within the context of a theoretical framework, the findings may have broader significance and utility. Researchers should have a conceptual rationale and conceptual definitions of key variables.

Step 5: Formulating Hypotheses

A **hypothesis** is a statement of the researcher's expectations or predictions about relationships among study variables. The research question identifies the study concepts and asks how the concepts might be related; a hypothesis is the predicted answer. For example, the research question might be: Is preeclamptic toxemia related to stress during pregnancy? This might be translated into the following hypothesis: Women with high levels of stress during pregnancy will be more likely than women with lower stress to experience preeclamptic toxemia. Most quantitative studies are designed to test hypotheses through statistical analysis.

Phase 2: The Design and Planning Phase

In the second major phase of a quantitative study, researchers make decisions about the methods they will use to address the research question. Researchers usually have considerable flexibility in designing a study, and they make many decisions. These methodologic decisions have crucial implications for the integrity of the resulting evidence. If the methods used to collect and analyze research data are flawed, then the evidence from the study may have little value.

Step 6: Selecting a Research Design

The **research design** is the overall plan for obtaining answers to the research questions. Many experimental and nonexperimental research designs are available. In designing the study, researchers select a specific design and identify strategies to minimize bias. Research designs indicate how often data will be collected, what types of comparisons will be made, and where the study will take place. The research design is the architectural backbone of the study.

Step 7: Developing Protocols for the Intervention

In experimental research, researchers actively intervene, which means that participants are exposed to different treatment conditions. For example, if we were interested in testing the effect of biofeedback in treating hypertension, the independent variable would be biofeedback compared with either an alternative treatment (e.g., relaxation), or no treatment. An **intervention protocol** for the study must be developed, specifying exactly what the biofeedback treatment would entail (e.g., who would administer it, how frequently, over how long a period the treatment would last, and so on) *and* what the alternative condition would be. The goal of well-articulated protocols is to have all people in each group treated in

the same way. (In nonexperimental research, this step is not necessary.)

Step 8: Identifying the Population to be Studied

Quantitative researchers need to clarify the group to whom study results can be generalized—that is, they must identify the population to be studied. A **population** is *all* the individuals or objects with common, defining characteristics. For example, the population of interest might be all patients undergoing chemotherapy in San Diego.

Step 9: Designing the Sampling Plan

Researchers collect data from a sample, which is a subset of the population. Using samples is more practical than collecting data from an entire population, but the risk is that the sample might not reflect the population's traits. In a quantitative study, a sample's adequacy is assessed by its size and **representativeness**. The quality of the sample depends on how typical, or representative, the sample is of the population. The **sampling plan** specifies how the sample will be selected and recruited, and how many subjects there will be.

Step 10: Specifying Methods to Measure Research Variables

Quantitative researchers must develop or borrow methods to measure the research variables accurately. Based on the conceptual definitions, researchers identify appropriate methods to operationalize variables and collect the data. The primary methods of data collection are *self-reports* (e.g., interviews), *observations* (e.g., observing the sleep–wake state of infants), and *biophysiologic measurements*. Measuring research variables and developing a **data collection plan** are challenging activities.

Step 11: Developing Methods to Safeguard Human/Animal Rights

Most nursing research involves humans, and so procedures need to be developed to ensure that the study adheres to ethical principles. Each aspect of the study plan needs to be scrutinized to determine whether the rights of participants have been adequately protected. A formal presentation to an ethics committee is often required.

Step 12: Reviewing and Finalizing the Research Plan

Before collecting their data, researchers often take steps to ensure that plans will work smoothly. For example, they may evaluate the *readability* of written materials to determine if participants with low reading skills can comprehend them, or they may *pretest* their measuring instruments to see if they work well. Normally, researchers also have their research plan critiqued by peers, consultants, or other reviewers before implementing it. Researchers seeking financial support submit a **proposal** to a funding source, and reviewers usually suggest improvements.

Phase 3: The Empirical Phase

The empirical phase of quantitative studies involves collecting data and preparing the data for analysis. Often, the empirical phase is the most time-consuming part of the investigation. Data collection typically requires many weeks, or even months, of work.

Step 13: Collecting the Data

The actual collection of data in quantitative studies often proceeds according to a preestablished plan. The plan specifies where and when the data will be gathered, procedures for describing the study to participants, and methods for recording information. Technological advances have expanded possibilities for automating data collection.

Step 14: Preparing the Data for Analysis

Data collected in a quantitative study are rarely amenable to direct analysis—preliminary steps are needed. One such step is **coding**, which is the process of translating verbal data into numeric form. For example, patients' responses to a question about their gender might be coded "1" for female and "2" for male (or vice versa). Another preliminary step involves entering the data onto computer files for analysis.

Phase 4: The Analytic Phase

Quantitative data are not reported in *raw* form (i.e., as a mass of numbers). They are subjected to

analysis and interpretation, which occurs in the fourth major phase of a project.

Step 15: Analyzing the Data

Quantitative researchers analyze their data through **statistical analyses**, which include simple procedures (e.g., computing an average) as well as ones that are complex. Some analytic methods are computationally formidable, but the underlying logic of statistical tests is fairly easy to grasp. Computers have eliminated the need to get bogged down with mathematic operations.

Step 16: Interpreting the Results

Interpretation involves making sense of study results and examining their implications. Researchers attempt to explain the findings in light of prior evidence, theory, and their own clinical experience—and in light of the adequacy of the methods, they used in the study. Interpretation also involves envisioning how the new evidence can best be used in clinical practice, and what further research is needed.

Phase 5: The Dissemination Phase

In the analytic phase, the researcher comes full circle: questions posed at the outset are answered. Researchers' responsibilities are not completed, however, until study results are disseminated.

Step 17: Communicating the Findings

A study cannot contribute evidence to nursing practice if the results are not shared. Another—and often final—task of a study, therefore, is the preparation of a **research report** that summarizes the study. Research reports can take various forms: dissertations, journal articles, conference presentations, and so on. Journal articles—reports appearing in such professional journals as *Nursing Research*—usually are the most useful because they are available to a broad, international audience. We discuss journal articles later in this chapter.

Step 18: Utilizing the Findings in Practice

Ideally, the concluding step of a high-quality study is to plan for the use of the evidence in practice settings. Although nurse researchers may not themselves be able to implement a plan for using the evidence, they can contribute to the process by including in their research reports recommendations regarding how the study evidence could be used in practice, by ensuring that adequate information has been provided for a meta-analysis, and by pursuing opportunities to disseminate the findings to clinicians.

ACTIVITIES IN A QUALITATIVE STUDY

Quantitative research involves a fairly linear progression of tasks—researchers plan the steps to be taken to maximize study integrity and then follow those steps as faithfully as possible. In qualitative studies, by contrast, the progression is closer to a circle than to a straight line—qualitative researchers are continually examining and interpreting data and making decisions about how to proceed based on what has already been discovered (Figure 3.2).

Because qualitative researchers have a flexible approach, it is impossible to define the flow of activities in a study precisely—the flow varies from one study to another, and researchers themselves do not know ahead of time exactly how the study will proceed. We try to provide a sense of how qualitative studies are conducted, however, by describing some major activities and indicating how and when they might be performed.

Conceptualizing and Planning a Qualitative Study

Identifying the Research Problem

Qualitative researchers usually begin with a broad topic area, focusing on an aspect of a topic that is poorly understood and about which little is known. They may not pose refined research questions at the outset. The general topic area may be narrowed and clarified on the basis of self-reflection and discussion with others, but researchers may proceed

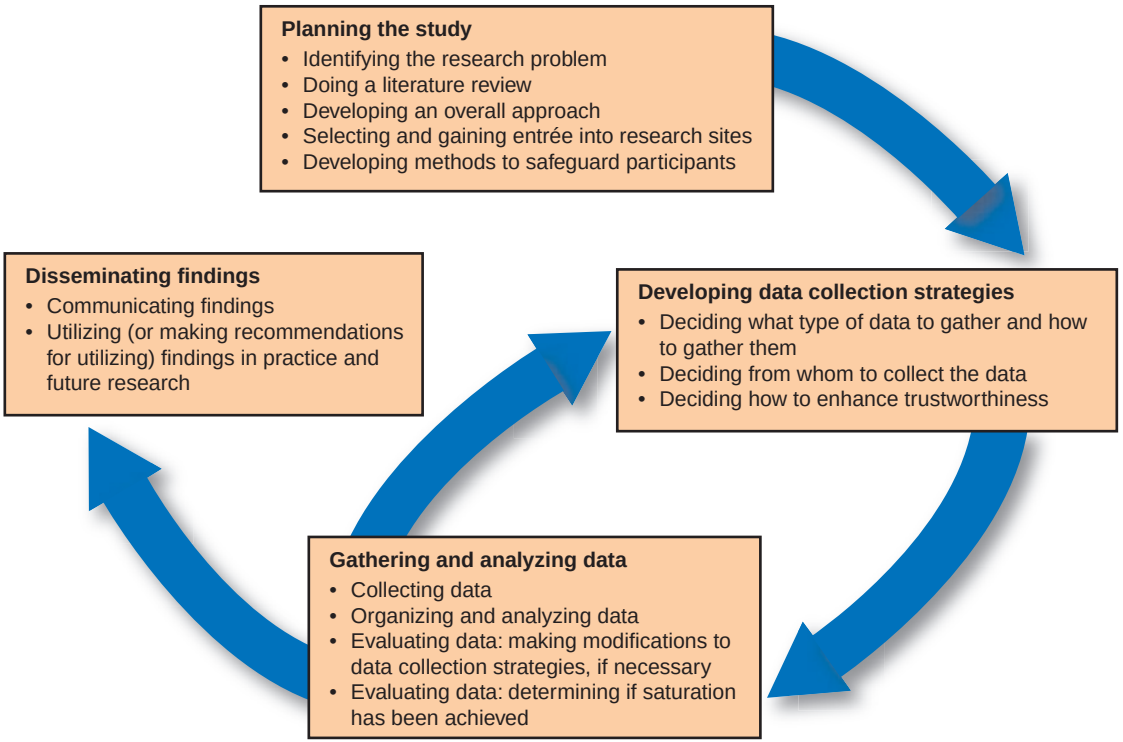


FIGURE 3.2 Flow of activities in a qualitative study.

initially with a fairly broad research question that allows the focus to be delineated more clearly, once the study is underway.

Doing a Literature Review

Qualitative researchers do not all agree about the value of an upfront literature review. Some believe that researchers should not consult the literature before collecting data, because prior studies could influence conceptualization of the focal phenomenon. In this view, the phenomena should be explicated based on participants' viewpoints rather than on prior knowledge. Those sharing this opinion often do a literature review at the end of the study. Other researchers conduct a brief preliminary review to get a general grounding. Still others believe that a full early literature review is appropriate. In any case, qualitative researchers typically find a fairly small body of relevant previous work because of the types of question they ask.

Selecting and Gaining Entrée into Research Sites

Before going into the field, qualitative researchers must identify an appropriate site. For example, if the topic is the health beliefs of the urban poor, an inner-city neighborhood with low-income residents must be identified. Researchers may need to engage in anticipatory fieldwork to identify a suitable and information-rich environment for the study. In some cases, researchers have ready access to the study site, but in others, they need to **gain entrée**. A site may be well suited to the needs of the research, but if researchers cannot "get in," the study cannot proceed. Gaining entrée typically involves negotiations with **gatekeepers** who have the authority to permit entry into their world.

Developing an Overall Approach in Qualitative Studies

Quantitative researchers do not collect data until the research design has been finalized. Qualitative researchers, by contrast, use an **emergent design** that materializes during the course of data collection. Certain design features may be guided by the qualitative research tradition within which the researcher is working, but nevertheless, few qualitative studies adopt rigidly structured designs that prohibit changes while in the field.

Although qualitative researchers do not always know in advance exactly how the study will progress, they nevertheless must have some sense of how much time is available for fieldwork and must also arrange for and test needed equipment, such as tape recorders or laptop computers. Other planning activities include such tasks as hiring and training interviewers to assist in the collection of data, securing interpreters if the informants speak a different language, and hiring appropriate consultants, transcribers, and support staff.

Addressing Ethical Issues

Qualitative researchers, like quantitative researchers, must also develop plans for addressing ethical issues—and, indeed, there are special concerns in qualitative studies because of the more intimate nature of the relationship that typically develops between researchers and study participants. Chapter 7 describes these concerns.

Conducting a Qualitative Study

In qualitative studies, the tasks of sampling, data collection, data analysis, and interpretation typically take place iteratively. Qualitative researchers begin by talking with or observing a few people with first-hand experience with the focal phenomenon. The discussions and observations are loosely structured, allowing for the expression of a full range of beliefs, feelings, and behaviors. Analysis and interpretation are ongoing, concurrent activities that guide choices about the kinds of people to sample next and the types of questions to ask or observations to make.

Data analysis involves clustering together related types of narrative information into a coherent scheme. As analysis and interpretation progress, researchers begin to identify **themes** and categories, which are used to build a rich description or theory of the phenomenon. The kinds of data obtained and the people selected as participants tend to become increasingly purposeful as the conceptualization is developed and refined. Concept development and verification shape the sampling process—as a conceptualization or theory develops, the researcher seeks participants who can confirm and enrich the theoretical understandings, as well as participants who can potentially challenge them and lead to further theoretical development.

Quantitative researchers decide upfront how many people to include in a study, but qualitative researchers' sampling decisions are guided by the data. Qualitative researchers use the principle of **data saturation**, which occurs when themes and categories in the data become repetitive and redundant, such that no new information can be gleaned by further data collection.

Quantitative researchers seek to collect high-quality data by using measuring instruments that have been demonstrated to be accurate and valid. Qualitative researchers, by contrast, must take steps to demonstrate the *trustworthiness* of the data while in the field. The central feature of these efforts is to confirm that the findings accurately reflect the experiences and viewpoints of participants. One confirmatory activity, for example, involves going back to participants and sharing preliminary interpretations with them so that they can evaluate whether the researcher's thematic analysis is consistent with their experiences.

Qualitative researchers sometimes need to develop appropriate strategies for leaving the field. Because qualitative researchers may develop strong relationships with participants and entire communities, they need to be sensitive to the fact that their departure might seem like a form of abandonment. Graceful departures and methods of achieving closure are important.

Disseminating Qualitative Findings

Qualitative nursing researchers also strive to share their findings with others at conferences and in journal articles. Qualitative findings, because of their depth and richness, also lend themselves to book-length manuscripts. Regardless of researchers' positions about *when* a literature review should be conducted, they usually include a summary of prior research in their reports as a means of providing context for the study.

Quantitative reports almost never contain **raw data**—that is, data in the form they were collected, which are numeric values. Qualitative reports, by contrast, are usually filled with rich verbatim passages directly from participants. The excerpts are used in an evidentiary fashion to support or illustrate researchers' interpretations and thematic construction.

Example of raw data in a qualitative report: Langegard and Ahlberg (2009) explored things that patients with incurable cancer had found consoling during the course of the disease. In-depth interviews with 10 hospice patients revealed that a major theme was acceptance, as illustrated by the following quote:

"Talking about it is a way of getting the truth into my head. Through putting my situation into words, it becomes a way of understanding and then I have a possibility to be consoled. If I don't understand the consequences of my disease, I can't possibly be consoled ... It's not about giving up, but it's about realizing that this is the way it is. It's over, it's incurable" (p. 104).

Like quantitative researchers, qualitative nurse researchers want their findings used by others. Qualitative findings often are the basis for formulating hypotheses that are tested by quantitative researchers, for developing measuring instruments for both research and clinical purposes, and for designing effective nursing interventions. Qualitative studies help to shape nurses' perceptions of a problem or situation, their conceptualizations of potential solutions, and their understanding of patients' concerns and experiences.

RESEARCH JOURNAL ARTICLES

Research **journal articles**, which summarize the context, design, and results of a study, are the primary method of disseminating research evidence. This section reviews the content and style of research journal articles to ensure that you will be equipped to delve into the research literature. A more detailed discussion of the structure of journal articles is presented in Chapter 28, which provides guidance on writing research reports.

Content of Journal Articles

Many quantitative and qualitative journal articles follow a conventional organization called the **IMRAD format**. This format, which loosely follows the steps of quantitative studies, involves organizing material into four main sections—**I**ntroduction, **M**ethod, **R**esults, **a**nd **D**iscussion. The main text of the report is usually preceded by an abstract and followed by references.

The Abstract

The **abstract** is a brief description of the study placed at the beginning of the article. The abstract answers, in about 200 words, the following: What were the research questions? What methods did the researcher use to address the questions? What did the researcher find? What are the implications for nursing practice? Readers can review an abstract to assess whether the entire report is of interest. Some journals have moved from traditional abstracts—single paragraphs summarizing the study's main features—to slightly longer, structured abstracts with specific headings. For example, abstracts in *Nursing Research* organize study information under the following headings: Background, Objectives, Method, Results, and Conclusions.

The Introduction

The introduction communicates the research problem and its context. The introduction, which often is not specifically labeled "Introduction," follows

immediately after the abstract. This section usually describes:

- The central phenomena, concepts, or variables under study
- The current state of evidence, based on a literature review
- The theoretical or conceptual framework
- The study purpose, research questions, or hypotheses to be tested
- The study's significance

Thus, the introduction sets the stage for a description of what the researcher did and what was learned. The introduction corresponds roughly to the conceptual phase (Phase 1) of a study.

The Method Section

The method section describes the methods used to answer the research questions. This section lays out methodologic decisions made in the design and planning phase (Phase 2), and may offer rationales for those decisions. In a quantitative study, the method section usually describes:

- The research design;
- The sampling plan;
- Methods of data collection and specific instruments used;
- Study procedures (including ethical safeguards); and
- Analytic procedures and methods.

Qualitative researchers discuss many of the same issues, but with different emphases. For example, a qualitative study often provides more information about the research setting and the study context, and less information on sampling. Also, because formal instruments are not used to collect qualitative data, there is less discussion about data collection methods, but there may be more information on data collection procedures. Increasingly, reports of qualitative studies are including descriptions of the researchers' efforts to enhance the rigor of the study.

The Results Section

The results section presents the **findings** (results) obtained in the data analyses. The text summarizes key findings, often accompanied by more detailed tables or figures. Virtually all results sections contain descriptive information, including a description of the participants (e.g., average age, percent male/female).

In quantitative studies, the results section provides information about **statistical tests**, which are used to test hypotheses and evaluate the believability of the findings. For example, if the percentage of smokers who smoke two packs or more daily is computed to be 40%, how *probable* is it that the percentage is accurate? If the researcher finds that the average number of cigarettes smoked weekly is lower for those in an intervention group than for those not getting the intervention, how *probable* is it that the intervention effect is *real*? Is the effect of the intervention on smoking likely to be replicated with a new sample of smokers—or does the result reflect a peculiarity of the sample? Statistical tests help to answer such questions. Researchers typically report:

- *The names of statistical tests used.* Different tests are appropriate for different situations, but they are based on common principles. You do not have to know the names of all statistical tests—there are dozens of them—to comprehend the findings.
- *The value of the calculated statistic.* Computers are used to calculate a numeric value for the particular statistical test used. The value allows researchers to draw conclusions about the meaning of the results. The *actual* numeric value of the statistic, however, is not inherently meaningful and need not concern you.
- *The significance.* A critical piece of information is whether the value of the statistic was significant (not to be confused with important or clinically relevant). When researchers report that results are **statistically significant**, it means the findings are probably reliable and replicable with a new sample. Research reports also indicate the **level of significance**, which is an index of how probable it is that the findings are reliable. For example, if a report says that a finding was significant at the .05 level, this means that

only 5 times out of 100 (5 ÷ 100 = .05) would the result be spurious. In other words, 95 times out of 100, similar results would be obtained with a new sample. Readers can have a high degree of confidence—but not total assurance—that the evidence is reliable.

Example from the results section of a quantitative study: Cook and colleagues (2009) studied degree of agreement between blood glucose values obtained by laboratory analysis versus by a point-of-care device. Their results indicated that, "Laboratory glucose values for blood from a catheter differed significantly from point-of-care values for blood from the catheter ($t = -9.18$, $p < .001$)" (p. 65). The average glucose value was 124 mg/dL for the point-of-care analysis, compared to 114 mg/dL for the laboratory analysis.

In this study, Cook and colleagues found that glucose values from the lab were significantly lower than those obtained from point-of-care devices. The average difference of 10 mg/dL was not likely to have been a haphazard difference, and would probably be replicated with a new sample. This finding is highly reliable: less than one time in 1,000 ($p < 0.001$) would a difference this great have occurred as a fluke. To understand this finding, you do not have to understand what a *t* statistic is, nor do you need to worry about the actual value of the statistic, −9.18.

Qualitative researchers often organize findings according to the major themes, processes, or categories identified in the data. Results sections of qualitative reports often have several subsections, the headings of which correspond to the themes. Excerpts from the raw data are presented to support and provide a rich description of the thematic analysis. The results section of qualitative studies may also present the researcher's emerging theory about the phenomenon under study.

The Discussion Section

In the discussion section, researchers draw conclusions about what the results mean, and how the evidence can be used in practice. The discussion often reviews study limitations and the implications of the limitations for the integrity of the results. Researchers are in the best position to point out sample deficiencies, design problems, weaknesses in data collection, and so forth. A discussion section that presents these limitations demonstrates to readers that the author was aware of these limitations and probably took them into account in interpreting the findings.

The Style of Research Journal Articles

Research reports tell a story. However, the style in which many research journal articles are written—especially reports of quantitative studies—makes it difficult for many readers to figure out or become interested in the story. To unaccustomed audiences, research reports may seem stuffy, pedantic, and bewildering. Four factors contribute to this impression:

1. *Compactness*. Journal space is limited, so authors compress a lot of information into a short space. Interesting, personalized aspects of the study cannot be reported; in qualitative studies, only a handful of supporting quotes can be included.
2. *Jargon*. The authors of research reports use terms that may seem esoteric.
3. *Objectivity*. Quantitative researchers tell their stories objectively, often in a way that makes them sound impersonal. For example, most quantitative reports are written in the passive voice (i.e., personal pronouns are avoided), which tends to make a report less inviting and lively than use of the active voice. Qualitative reports, by contrast, are more subjective and personal, and written in a more conversational style.
4. *Statistical information*. The majority of nursing studies are quantitative, and thus most reports summarize the results of statistical analyses. Numbers and statistical symbols can intimidate readers who do not have statistical training.

In this textbook, we try to assist you in dealing with these issues and also strive to encourage you to tell *your* research stories in a manner that makes them accessible to practicing nurses.

Tips on Reading Research Reports

As you progress through this textbook, you will acquire skills for evaluating various aspects of research reports critically. Some preliminary hints on digesting research reports follow.

- Grow accustomed to the style of research articles by reading them frequently, even though you may not yet understand all the technical points.
- Read from an article that has been copied (or downloaded and printed) so that you can highlight portions and write marginal notes.
- Read articles slowly. Skim the article first to get major points and then read it more carefully a second time.
- On the second reading of a journal article, train yourself to be an *active* reader. Reading actively means that you constantly monitor yourself to assess your understanding of what you are reading. If you have problems, go back and reread difficult passages or make notes so that you can ask someone for clarification. In most cases, that "someone" will be your research instructor, but also consider contacting researchers themselves via e-mail.
- Keep this textbook with you as a reference while you are reading articles so that you can look up unfamiliar terms in the glossary or index.
- Try not to get bogged down in (or scared away by) statistical information. Try to grasp the gist of the story without letting numbers frustrate you.
- Until you become accustomed to research journal articles, you may want to "translate" them by expanding compact paragraphs into looser constructions, by translating jargon into familiar terms, by recasting the report into an active voice, and by summarizing findings with words rather than numbers. (Chapter 3 in the accompanying *Resource Manual* has an example of such a translation).

GENERAL QUESTIONS IN REVIEWING A RESEARCH STUDY

Most chapters of this book contain guidelines to help you evaluate different aspects of a research report critically, focusing primarily on the researchers' methodologic decisions. Box 3.3 presents some further suggestions for performing a preliminary overview of a research report, drawing on concepts explained in this chapter. These guidelines supplement those presented in Box 1.1, Chapter 1.

BOX 3.3 Additional Questions for a Preliminary Review of a Study

1. What is the study all about? What are the main phenomena, concepts, or constructs under investigation?
2. If the study is quantitative, what are the independent and dependent variables?
3. Do the researchers examine relationships or patterns of association among variables or concepts? Does the report imply the possibility of a causal relationship?
4. Are key concepts clearly defined, both conceptually and operationally?
5. What type of study does it appear to be, in terms of types described in this chapter: Quantitative—experimental? nonexperimental? Qualitative—descriptive? grounded theory? phenomenology? ethnography?
6. Does the report provide any information to suggest how long the study took to complete?
7. Does the format of the report conform to the traditional IMRAD format? If not, in what ways does it differ?

RESEARCH EXAMPLES

In this section, we illustrate the progression of activities and discuss the time schedule of two studies (one quantitative and the other qualitative) conducted by the second author of this book.

Project Schedule for a Quantitative Study

Beck and Gable (2001) undertook a study to evaluate a scale they developed, the Postpartum Depression Screening Scale (PDSS).

Phase 1. Conceptual Phase: 1 Month

This phase was short, because much of the conceptual work had been done in an earlier study, in which Beck and Gable developed the PDSS. The literature had already been reviewed and Beck had done extensive fieldwork. The same framework and conceptual definitions that had been used in the first study were used in the new study.

Phase 2. Design and Planning Phase: 6 Months

The second phase included fine tuning the research design, gaining entrée into the hospital where subjects were recruited, and obtaining approval of the hospital's human subjects review committee. During this period, Beck met with statistical consultants and with Gable, an instrument development specialist, numerous times.

Phase 3. Empirical Phase: 11 Months

Data collection took almost a year to complete. The design called for administering the PDSS to 150 mothers at 6 weeks postpartum, and scheduling them for a psychiatric diagnostic interview to determine if they were suffering from postpartum depression. Recruitment of the women, which occurred in prepared childbirth classes, began 4 months before data collection. The researchers then waited until 6 weeks after delivery to gather data. The nurse psychotherapist, who had her own clinical practice, was able to come to the hospital only 1 day a week to conduct the diagnostic interviews; this contributed to the time required to achieve the desired sample size.

Phase 4. Analytic Phase: 3 Months

Statistical tests were performed to determine a cutoff score on the PDSS above which mothers would be identified as having screened positive for postpartum depression. Data analysis also was undertaken to determine the accuracy of the PDSS in predicting diagnosed postpartum depression. During this phase, Beck met with Gable and statisticians to interpret results.

Phase 5. Dissemination Phase: 18 Months

The researchers prepared and submitted their report to the journal *Nursing Research* for possible publication. It was accepted within 4 months, but it was "in press" (awaiting publication) for 14 months before being published. During this period, the authors presented their findings at regional and international conferences.

Project Schedule for a Qualitative Study

Beck (2004) conducted a phenomenological study on women's experiences of birth trauma. Total time from start to finish was approximately 3 years.

Phase 1. Conceptual Phase: 3 Months

Beck, who is renowned for her program of research on postpartum depression, became interested in birth trauma when she delivered the keynote address at a conference in New Zealand. She was asked to speak on perinatal anxiety disorders. In preparing for her address, Beck located only a handful of articles on birth trauma and its resulting post-traumatic stress disorder (PTSD). Following her keynote speech, a mother made a riveting presentation about her experience of

PTSD due to a traumatic childbirth. The mother, Sue Watson, was one of the founders of Trauma and Birth Stress (TABS), a charitable trust in New Zealand. Watson and Beck discussed the possibility of Beck conducting a qualitative study with the mothers who were members of TABS. Gaining entrée into TABS was facilitated by Watson and four other founders of TABS.

Phase 2. Design and Planning Phase: 3 Months

Beck selected a phenomenological design to describe the experience of a traumatic birth. Beck and Watson decided that Beck would write an introductory letter explaining the study, and Watson would write a letter endorsing the study. Both letters were to be sent to mothers who were members of TABS, asking for their cooperation. Once the basic design was developed, the research proposal was submitted to and approved by the ethics committee at Beck's university.

Phase 3. Empirical/Analytic Phases: 24 months

Data for the study were collected over an 18-month period, during which 40 mothers sent their stories of birth trauma to Beck via e-mail attachments. For the next 6 months, Beck analyzed the mothers' stories. Four themes emerged from data analysis: To care for me: Was that too much to ask? To communicate with me: Why was this neglected? To provide safe care: You betrayed my trust and I felt powerless, and The end justifies the means: At whose expense, at what price?

Phase 4 Dissemination Phase: 9 Months

A manuscript describing this study was submitted for publication to *Nursing Research* in April 2003. In June, Beck received a letter indicating that the reviewers' recommended she revise and resubmit the paper. Six weeks later, Beck resubmitted her revised manuscript, and in September, she was notified that her revised manuscript had been accepted for publication. The article was published in the January/February 2004 issue. Beck also has presented the findings at numerous national and international research conferences.

SUMMARY POINTS

- The people who provide information to the **researchers** (investigators) in a study are called **subjects** or **study participants** (in quantitative research) or study participants or **informants** in qualitative research; collectively they comprise the **sample**.
- The *site* is the overall location for the research; researchers sometimes engage in **multisite studies**. *Settings* are the more specific places where data collection occurs. Settings can range from totally naturalistic environments to formal laboratories.
- Researchers investigate **concepts** and **phenomena** (or **constructs**), which are abstractions or mental representations inferred from behavior or characteristics.
- Concepts are the building blocks of **theories**, which are systematic explanations of some aspect of the real world.
- In quantitative studies, concepts are called *variables*. A **variable** is a characteristic or quality that takes on different values (i.e., varies from one person to another). Groups that are varied with respect to an attribute are **heterogeneous**; groups with limited variability are **homogeneous**.
- **Continuous variables** can take on an infinite range of values along a continuum (e.g., weight). **Discrete variables** have a finite number of values between two points (e.g., number of children). **Categorical variables** have distinct categories that do not represent a quantity (e.g., gender).
- The **dependent** (or **outcome**) variable is the behavior or characteristic the researcher is interested in explaining, predicting, or affecting. The **independent variable** is the presumed cause of, antecedent to, or influence on the dependent variable.

- A **conceptual definition** describes the abstract or theoretical meaning of a concept being studied. An **operational definition** specifies procedures required to measure a variable.
- **Data**—information collected during a study—may take the form of narrative information (**qualitative data**) or numeric values (**quantitative data**).
- A **relationship** is a bond or connection between two variables. Quantitative researchers examine the relationship between the independent variable and dependent variable.
- When the independent variable causes or affects the dependent variable, the relationship is a **cause-and-effect** (or **causal**) **relationship**. In a **functional (associative) relationship**, variables are related in a noncausal way.
- A basic distinction in quantitative studies is between **experimental research**, in which researchers actively intervene, and **nonexperimental** (or **observational**) **research**, in which researchers make observations of existing phenomena without intervening.
- Qualitative research sometimes is rooted in research traditions that originate in other disciplines. Three such traditions are grounded theory, phenomenology, and ethnography.
- **Grounded theory** seeks to describe and understand key social psychological processes that occur in a social setting.
- **Phenomenology** focuses on the lived experiences of humans and is an approach to learning what the life experiences of people are like and what they mean.
- **Ethnography** provides a framework for studying the meanings and lifeways of a culture in a holistic fashion.
- Quantitative researchers usually progress in a fairly linear fashion from asking research questions to answering them. The main phases in a quantitative study are the conceptual, planning, empirical, analytic, and dissemination phases.
- The *conceptual phase* involves (1) defining the problem to be studied, (2) doing a **literature review**, (3) engaging in **clinical fieldwork** for clinical studies, (4) developing a framework and conceptual definitions, and (5) formulating **hypotheses** to be tested.
- The *planning phase* entails (6) selecting a **research design**, (7) developing **intervention protocols** if the study is experimental, (8) specifying the **population**, (9) developing a **sampling plan**, (10) specifying methods to measure the research variables, (11) developing strategies to safeguard the rights of participants, and (12) finalizing the research plan (e.g., *pretesting* instruments).
- The *empirical phase* involves (13) collecting data and (14) preparing data for analysis.
- The *analytic phase* involves (15) analyzing data through **statistical analysis** and (16) interpreting the results.
- The *dissemination phase* entails (17) communicating the findings in a **research report** and (18) promoting the use of the study evidence in nursing practice.
- The flow of activities in a qualitative study is more flexible and less linear. Qualitative studies typically involve an **emergent design** that evolves during **fieldwork**.
- Qualitative researchers begin with a broad question regarding a phenomenon, often focusing on a little-studied aspect. In the early phase of a qualitative study, researchers select a site and seek to **gain entrée** into it, which typically involves enlisting the cooperation of **gatekeepers**.
- Once in the field, researchers select informants, collect data, and then analyze and interpret them in an iterative fashion; field experiences help in an ongoing fashion to shape the design of the study.
- Early analysis in qualitative research leads to refinements in sampling and data collection, until **saturation** (redundancy of information) is achieved.
- Both qualitative and quantitative researchers disseminate their findings, most often in **journal articles** that concisely communicate what the researchers did and what they found.

- Journal articles often consist of an **abstract** (a brief synopsis) and four major sections in an **IMRAD format**: an **I**ntroduction (explanation of the study problem and its context), **M**ethod section (the strategies used to address the problem), **R**esults section (study findings), **a**nd **D**iscussion (interpretation of the findings).
- Research reports are often difficult to read because they are dense and contain a lot of jargon. Quantitative research reports may be intimidating at first because, compared to qualitative reports, they are more impersonal and report on statistical tests.
- **Statistical tests** are procedures for testing research hypotheses and evaluating the believability of the findings. Findings that are **statistically significant** are ones that have a high probability of being "real."

STUDY ACTIVITIES

Chapter 3 of the *Resource Manual for Nursing Research: Generating and Assessing Evidence for Nursing Practice, 9th ed.*, offers study suggestions for reinforcing concepts presented in this chapter. In addition, the following questions can be addressed in classroom or online discussions:

1. Suggest ways of conceptually and operationally defining the following concepts: nursing competency, aggressive behavior, pain, postsurgical recovery, and body image.
2. Name five continuous, five discrete, and five categorical variables and identify which, if any, are dichotomous.
3. In the following research problems, identify the independent and dependent variables:
   a. Does screening for intimate partner violence among pregnant women improve birth and delivery outcomes?
   b. Do elderly patients have lower pain thresholds than younger patients?
   c. Are the sleeping patterns of infants affected by different forms of stimulation?
   d. Can home visits by nurses to released psychiatric patients reduce readmission rates?

STUDIES CITED IN CHAPTER 3

Al-Akour, N., Khader, Y., & Shatnawi, N. (2010). Quality of life and associated factors among *Jordanian* adolescents with type 1 diabetes mellitus. *Journal of Diabetes and its Complications, 24*, 43–47.

Beck, C. T. (2004). Birth trauma: In the eye of the beholder. *Nursing Research, 53*, 28–35.

Beck, C. T. (2009). The arm: There is no escaping the reality for mothers of children with obstetric brachial plexus injuries. *Nursing Research, 58*, 237–245.

Beck, C. T., & Gable, R. K. (2001). Further validation of the Postpartum Depression Screening Scale. *Nursing Research, 50*, 155–164.

Cook, S., Laughlin, D., Moore, M., North, D., Wilkins, K., Wong, G., Wallace-Scroggs, A., & Halvorson, L. (2009). Differences in glucose values obtained from point-of-care glucose meters and laboratory analysis in critically ill patients. *American Journal of Critical Care, 18*, 65–71.

Gaudine, A., Gien, L., Thuan, T., & Dung-do, V. (2010). Perspectives of HIV-related stigma in a community in Vietnam. *International Journal of Nursing Studies, 47*, 38–48.

Hessler, K. (2009). Physical activity behaviors of rural preschoolers. *Pediatric Nursing, 35*, 246–253.

Langegard, U., & Ahlberg, K. (2009). Consolation in conjunction with incurable cancer. *Oncology Nursing Forum, 36*, 99–107.

Lin, H., Tsai, Y., Lin, P., & Tsay, P. (2010). Effects of a therapeutic lifestyle change programme on cardiac risk factors after coronary artery bypass graft. *Journal of Clinical Nursing, 19*, 60–68.

Marshall, J., Cowell, J., Campbell, E., & McNaughton, D. (2010). Regional variations in cancer screening rates found in women with diabetes. *Nursing Research, 59*, 34–41.

Polit, D. F., London, A., & Martinez, J. (2001). *The health of poor urban women.* New York: Manpower Demonstration Research Corporation. (Report available online at: www.mdrc.org)

Propp, K., Apker, J., Ford, W., Wallace, N., Serbenski, M., & Hofmeister, N. (2010). Meeting the complex needs of the health care team: Identification of nurse-team communication practices perceived to enhance patient outcomes. *Qualitative Health Research, 20*, 15–28.

Schachman, K. (2010). Online fatherhood: The experience of first-time fatherhood in combat-deployed troops. *Nursing Research, 59*, 11–17.

Schim, S., Doorenbos, A., & Borse, N. (2006) Enhancing cultural competence among hospice staff. *The American Journal of Hospice & Palliative Care, 23,* 404–411.

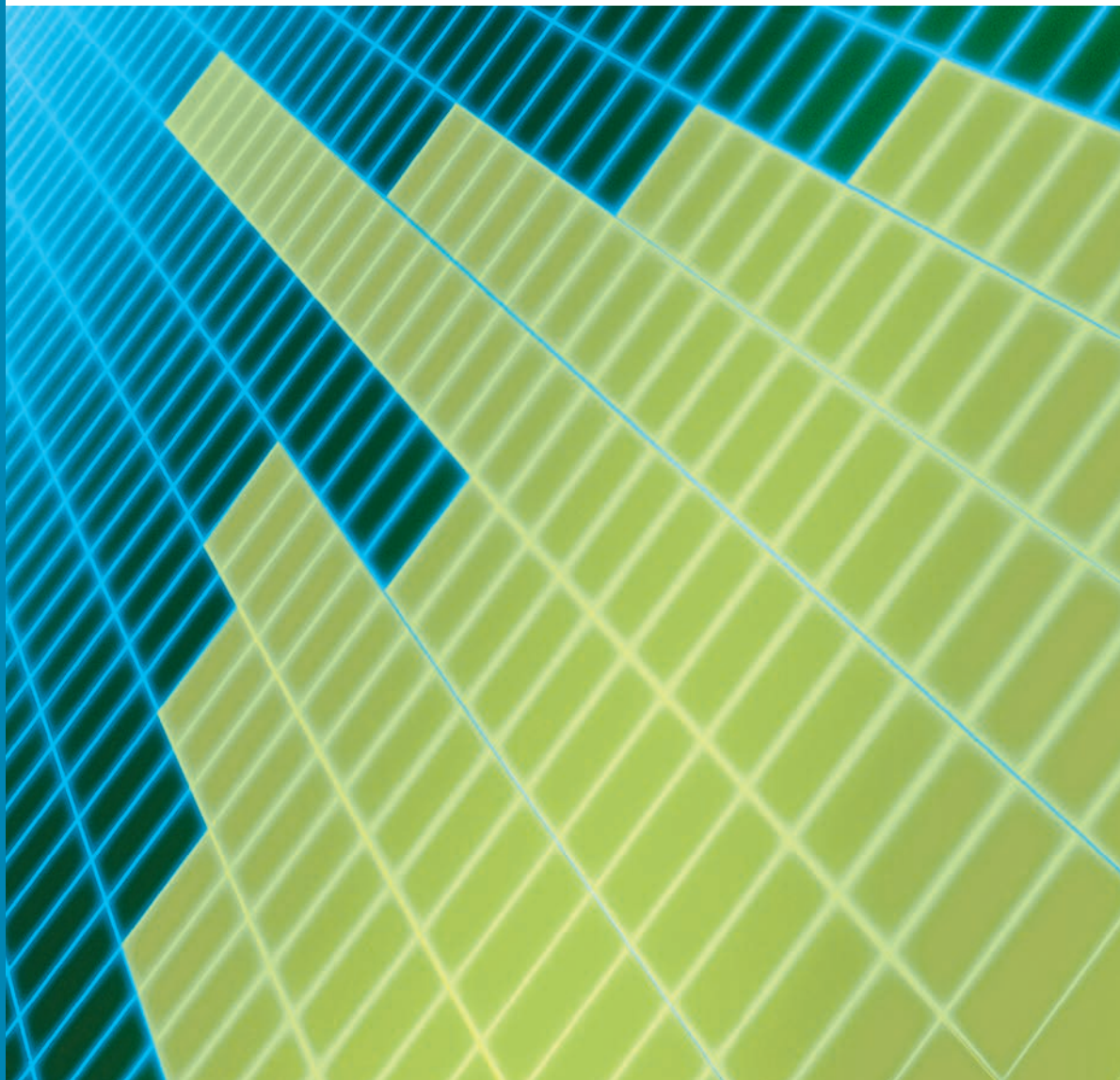
Twiss, J., Waltman, N., Berg, K., Ott, C., Gross, G., & Lindsey, A. (2009). An exercise intervention for breast cancer survivors with bone loss. *Journal of Nursing Scholarship, 41*, 20–27.

Vallance, J., Plotnikoff, R., Karvinen, K., Mackey, J., & Courneya, K. (2010). Understanding physical activity maintenance in breast cancer survivors. *American Journal of Health Behavior, 34*, 225–236.

Methodologic and nonresearch references cited in this chapter can be found in a separate section at the end of the book.

PART 2

CONCEPTUALIZING AND PLANNING A STUDY TO GENERATE EVIDENCE FOR NURSING



4 Research Problems, Research Questions, and Hypotheses

OVERVIEW OF RESEARCH PROBLEMS

Studies begin much like an EBP effort—as problems that need to be solved, or as questions that need to be answered. This chapter discusses the development of research problems. We begin by clarifying some relevant terms.

Basic Terminology

At a general level, a researcher selects a **topic** or a phenomenon on which to focus. Examples of research topics are claustrophobia during MRI tests, pain management for sickle cell disease, and nutrition during pregnancy. Within these broad topic areas are many potential research problems. In this section, we illustrate various terms using the topic *side effects of chemotherapy.*

A **research problem** is an enigmatic or troubling condition. Researchers identify a research problem within a broad topic area of interest. The purpose of research is to "solve" the problem—or to contribute to its solution—by generating relevant evidence. A **problem statement** articulates the problem and describes the need for a study through the development of an *argument*. Table 4.1 presents a simplified problem statement related to the topic of side effects of chemotherapy.

Research questions are the specific queries researchers want to answer in addressing the problem. Research questions guide the types of data to collect in a study. Researchers who make specific predictions about answers to research questions pose **hypotheses** that are then tested.

Many reports include a **statement of purpose** (or purpose statement), which summarizes the study goals. Researchers might also identify several **research aims** or **objectives**—the specific accomplishments they hope to achieve by conducting the study. The objectives include answering research questions or testing research hypotheses, but may also encompass broader aims (e.g., developing an effective intervention).

These terms are not always consistently defined in research methods textbooks, and differences among them are often subtle. Table 4.1 illustrates the interrelationships among terms as we define them.

Research Problems and Paradigms

Some research problems are better suited to qualitative versus quantitative methods. Quantitative studies usually focus on concepts that are fairly well developed, about which there is an existing body of evidence, and for which there are reliable methods of measurement. For example, a quantitative study might be undertaken to explore whether

TABLE 4.1 Example of Terms Relating to Research Problems

TERM	EXAMPLE
Topic/focus	Side effects of chemotherapy
Research problem (Problem statement)	Nausea and vomiting are common side effects among patients on chemotherapy, and interventions to date have been only moderately successful in reducing these effects. New interventions that can reduce or prevent these side effects need to be identified.
Statement of purpose	The purpose of the study is to test an intervention to reduce chemotherapy-induced side effects—specifically, to compare the effectiveness of patient-controlled and nurse-administered antiemetic therapy for controlling nausea and vomiting in patients on chemotherapy.
Research question	What is the relative effectiveness of patient-controlled antiemetic therapy versus nurse-controlled antiemetic therapy with regard to (a) medication consumption and (b) control of nausea and vomiting in patients on chemotherapy?
Hypotheses	Subjects receiving antiemetic therapy by a patient-controlled pump will (1) be less nauseous, (2) vomit less, and (3) consume less medication than subjects receiving the therapy by nurse administration.
Aims/objectives	This study has as its aim the following objectives: (1) to develop and implement two alternative procedures for administering antiemetic therapy for patients receiving moderate emetogenic chemotherapy (patient controlled versus nurse controlled), (2) to test three hypotheses concerning the relative effectiveness of the alternative procedures on medication consumption and control of side effects, and (3) to use the findings to develop recommendations for possible changes to clinical procedures.

older people with chronic illness who continue working are less (or more) depressed than those who retire. There are relatively accurate measures of depression that would yield quantitative information about the level of depression in a sample of employed and retired chronically ill older people.

Qualitative studies are often undertaken because some aspect of a phenomenon is poorly understood, and the researcher wants to develop a rich and context-bound understanding of it. Qualitative studies are often initiated to heighten awareness and create a dialogue about a phenomenon. Qualitative methods would not be well suited to comparing levels of depression among employed and retired seniors, but they would be ideal for exploring, for example, the *meaning* of depression among chronically ill retirees. Thus, the nature of the research question is closely allied to paradigms and to research traditions within paradigms.

Sources of Research Problems

Where do ideas for research problems come from? At a basic level, research topics originate with

researchers' interests. Because research is a time-consuming enterprise, inquisitiveness about and interest in a topic are essential. Research reports rarely indicate the source of researchers' inspiration, but a variety of explicit sources can fuel their curiosity, including the following:

- *Clinical experience.* Nurses' everyday clinical experience is a rich source of ideas for research topics. Immediate problems that need a solution—analogous to problem-focused triggers discussed in Chapter 2—may generate more enthusiasm than abstract problems inferred from a theory, and they have high potential for clinical significance.
- *Quality improvement efforts.* Important clinical questions sometimes emerge in the context of findings from quality improvement studies. Personal involvement on a quality improvement team can sometimes generate ideas for a study.
- *Nursing literature.* Ideas for studies often come from reading the nursing literature. Research articles may suggest problems indirectly by stimulating the reader's curiosity and directly by identifying needed research. Familiarity with existing research or with emerging clinical issues is an important route to developing a research topic.
- *Social issues.* Topics are sometimes suggested by global social or political issues of relevance to the healthcare community. For example, the feminist movement raised questions about such topics as gender equity in healthcare. Public awareness about health disparities has led to research on healthcare access and culturally sensitive interventions.
- *Theories.* Theories from nursing and related disciplines are another source of research problems. Researchers ask, If this theory is correct, what would I predict about people's behaviors, states, or feelings? The predictions can then be tested through research.
- *Ideas from external sources.* External sources and direct suggestions can sometimes provide the impetus for a research idea. For example, ideas for studies may emerge by reviewing a funding agency's research priorities or from brainstorming with other nurses.

Additionally, researchers who have developed a *program of research* on a topic area may get inspiration for "next steps" from their own findings or from a discussion of those findings with others.

Example of a problem source for a quantitative study: Beck, one of this book's authors, has developed a strong research program on postpartum depression (PPD). Beck was approached by Dr. Carol Lammi-Keefe, a professor in nutritional sciences, who had been researching the effect of DHA (docosahexaemoic acid, a fat found in cold-water fish) on fetal brain development. The literature suggested that DHA might play a role in reducing the severity of PPD and so the two researchers are collaborating in a project to test the effectiveness of dietary supplements of DHA on the incidence and severity of PPD. Their clinical trial, funded by the Donaghue Medical Research Foundation, is currently underway.

TIP: Personal experiences in clinical settings are a provocative source of research ideas. Here are some hints on how to proceed:

- Watch for a recurring problem and see if you can discern a pattern in situations that lead to the problem.

Example: Why do many patients complain of being tired after being transferred from a coronary care unit to a progressive care unit?

- Think about aspects of your work that are frustrating or do not result in the intended outcome—then try to identify factors contributing to the problem that could be changed.

Example: Why is suppertime so frustrating in a nursing home?

- Critically examine your own clinical decisions. Are they based on tradition, or are they based on systematic evidence that supports their efficacy?

Example: What would happen if you used the return of flatus to assess the return of GI motility after abdominal surgery, rather than listening to bowel sounds?

DEVELOPING AND REFINING RESEARCH PROBLEMS

Unless a research problem is based on an explicit suggestion, actual procedures for developing one

are difficult to describe. The process is rarely a smooth and orderly one; there are likely to be false starts, inspirations, and setbacks. The few suggestions offered here are not intended to imply that there are techniques for making this first step easy but rather to encourage you to persevere in the absence of instant success.

Selecting a Topic

Developing a research problem is a creative process. In the early stages of generating research ideas, it is unwise to be too self-critical. It is better to relax and jot down areas of interest as they come to mind. It matters little if the terms you use to remind you of the ideas are abstract or concrete, broad or specific, technical or colloquial—the important point is to put ideas on paper.

After this first step, the ideas can be sorted in terms of interest, knowledge about the topics, and the perceived feasibility of turning the topics into a study. When the most fruitful idea has been selected, the list should not be discarded; it may be necessary to return to it.

TIP: The process of selecting and refining a research problem usually takes longer than you might think. The process involves starting with some preliminary ideas, having discussions with colleagues and advisers, persuing the research literature, looking at what is happening in clinical settings, and a lot of reflection.

Narrowing the Topic

Once you have identified a topic of interest, you can begin to ask some broad questions that can lead you to a researchable problem. Examples of question stems that may help to focus an inquiry include the following:

- What is going on with . . . ?
- What is the process by which . . . ?
- What is the meaning of . . . ?
- What is the extent of . . . ?
- What influences or causes . . . ?
- What differences exist between . . . ?
- What are the consequences of . . . ?
- What factors contribute to . . . ?

Again, early criticism of ideas can be counterproductive. Try not to jump to the conclusion that an idea sounds trivial or uninspired without giving it more careful consideration or exploring it with others.

Beginning researchers often develop problems that are too broad in scope or too complex for their level of methodologic expertise. The transformation of the general topic into a workable problem is typically accomplished in uneven steps. Each step should result in progress toward the goals of narrowing the scope of the problem and sharpening and defining the concepts.

As researchers move from general topics to more specific researchable problems, multiple potential problems can emerge. Consider the following example. Suppose you were working on a medical unit and were puzzled by the fact that some patients always complained about having to wait for pain medication when certain nurses were assigned to them. The general problem area is discrepancy in patient complaints regarding pain medications administered by different nurses. You might ask: What accounts for the discrepancy? How can I improve the situation? These queries are not research questions, but they may lead you to ask such questions as the following: How do the two groups of nurses differ? What characteristics do the complaining patients share? At this point, you may observe that the ethnic background of the patients and nurses could be relevant. This may lead you to search the literature for studies about ethnicity in relation to nursing care, or it may provoke you to discuss the observations with others. These efforts may result in several research questions, such as the following:

- What is the essence of patient complaints among patients of different ethnic backgrounds?
- Is the ethnic background of nurses related to the frequency with which they dispense pain medication?
- Does the number of patient complaints increase when patients are of dissimilar ethnic backgrounds as opposed to when they are of the same ethnic background as nurses?

- Do nurses' dispensing behaviors change as a function of the similarity between their own ethnic background and that of patients?

These questions stem from the same problem, yet each would be studied differently; for example, some suggest a qualitative approach and others suggest a quantitative one. A quantitative researcher might become curious about ethnic differences in nurses' dispensing behaviors. Both ethnicity and nurses' dispensing behaviors are variables that can be measured reliably. A qualitative researcher who noticed differences in patient complaints would likely be more interested in understanding the *essence* of the complaints, the patients' *experience* of frustration, or the *process* by which the problem got resolved. These are aspects of the research problem that would be difficult to quantify.

Researchers choose a problem to study based on several factors, including its inherent interest and its compatibility with a paradigm of preference. In addition, tentative problems vary in their feasibility and worth. A critical evaluation of ideas is appropriate at this point.

Evaluating Research Problems

There are no rules for making a final selection of a research problem, but some criteria should be kept in mind. Four important considerations are the problem's significance, researchability, feasibility, and interest to you.

Significance of the Problem

A crucial factor in selecting a problem is its significance to nursing. Evidence from the study should have potential to contribute meaningfully to nursing practice. Within the existing body of evidence, the new study should be the right "next step." The right next step could involve an original inquiry, but it could also be a replication to answer previously asked questions with greater rigor or with different types of people.

In evaluating the significance of an idea, the following kinds of questions are relevant: Is the problem important to nursing and its clients? Will patient care benefit from the evidence? Will the findings challenge (or lend support to) untested assumptions? If the answer to all these questions is "no," then the problem should be abandoned.

Researchability of the Problem

Not all problems are amenable to research inquiry. Questions of a moral or ethical nature, although provocative, cannot be researched. For example, should assisted suicide be legalized? There are no *right* or *wrong* answers to this question, only points of view. To be sure, it is possible to ask related questions that could be researched, such as the following:

- What are nurses' attitudes toward assisted suicide?
- What moral dilemmas are perceived by nurses who might be involved in assisted suicide?
- Do terminally ill patients living with high levels of pain hold more favorable attitudes toward assisted suicide than those with less pain?

The findings from studies addressing such questions would have no bearing on whether assisted suicide should be legalized, but the information could be useful in developing a better understanding of the issues.

Feasibility of Addressing the Problem

A third consideration concerns feasibility, which encompasses several issues. Not all of the following factors are universally relevant, but they should be kept in mind in making a decision.

Time. Most studies have deadlines or goals for completion, so the problem must be one that can be studied in the given time. The scope of the problem should be sufficiently restricted so that there will be enough time for the various steps reviewed in Chapter 3. It is prudent to be conservative in estimating time for various tasks because research activities often require more time than anticipated.

Availability of Study Participants. In any study involving humans, researchers need to consider whether people with the desired characteristics will be available and willing to cooperate. Securing people's cooperation is sometimes easy (e.g., getting nursing students to complete a questionnaire), but other situations pose more difficulties. Some people

may not have the time or interest, and others may not feel well enough to participate. If the research is time-consuming or demanding, researchers may need to exert extra effort in recruiting participants, or may have to offer a monetary incentive.

Cooperation of Others. It may be insufficient to get the cooperation of prospective participants alone. As noted in Chapter 3, it may be necessary to gain entrée into an appropriate community or setting, and to develop the trust of gatekeepers. In institutional settings (e.g., hospitals), access to clients, personnel, or records requires authorization. Most healthcare organizations require approval of proposed studies.

Facilities and Equipment. All studies have resource requirements, although needs are sometimes modest. It is prudent to consider what facilities and equipment will be needed and whether they will be available before embarking on a study. For example, if technical equipment is needed, can it be secured, and is it functioning properly? Availability of space, office equipment, and research support staff may also need to be considered.

Money. Monetary needs for studies vary widely, ranging from $100 to $200 for small student projects to hundreds of thousands of dollars for large-scale research. If you are on a limited budget, you should think carefully about projected expenses before selecting a problem. Major categories of research-related expenditures include:

- Personnel costs—payments to individuals hired to help with the study (e.g., for conducting interviews, coding, data entry, transcribing, word processing)
- Participant costs—payments to participants as an incentive for their cooperation or to offset their expenses (e.g., transportation or baby-sitting costs)
- Supplies—paper, envelopes, computer disks, postage, audiotapes, and so on
- Printing and duplication costs—expenses for reproducing forms, questionnaires, and so forth
- Equipment—laboratory apparatus, computers and software, audio or video recorders, calculators, and the like
- Laboratory fees for the analysis of biophysiologic data
- Transportation costs (e.g., travel to participants' homes)

Researcher Experience. The problem should be chosen from a field about which you have some prior knowledge or experience. Researchers may struggle with a topic that is new and unfamiliar—although upfront clinical fieldwork may make up for certain deficiencies. The issue of technical expertise also should be considered. Beginning researchers with limited methodologic skills should avoid research problems that might require the development of sophisticated measuring instruments or that involve complex analyses.

Ethical Considerations. A research problem may be unfeasible if an investigation of the problem would pose unfair or unethical demands on participants. An overview of major ethical considerations in research is presented in Chapter 7 and should be reviewed when considering the study's feasibility.

Researcher Interest

Even if a tentative problem is researchable, significant, and feasible, there is one more criterion: your own interest in the problem. Genuine fascination with the chosen research problem is an important prerequisite to a successful study. A lot of time and energy are expended in a study; there is little sense devoting these resources to a project about which you are not enthusiastic.

TIP: Beginning researchers often seek suggestions about a topic area, and such assistance may be helpful in getting started. Nevertheless, it is rarely wise to be talked into a topic toward which you are not personally inclined. If you do not find a problem attractive or stimulating during the beginning phases of a study, then you are bound to regret your choice later.

COMMUNICATING RESEARCH PROBLEMS

Every study needs a problem statement—an articulation of what it is that is problematic and that is the

impetus for the research. Most research reports also present either a statement of purpose, research questions, or hypotheses, and often combinations of these elements are included.

Many beginning researchers do not really understand problem statements and may even have trouble identifying them in a research article—not to mention developing one. A problem statement is presented early, and often begins with the very first sentence after the abstract. Specific research questions, purposes, or hypotheses appear later in the introduction. Typically, however, researchers *begin* their inquiry with a research question or a purpose, and *then* develop an argument in a problem statement to present the rationale for the new research. This section describes the wording of statements of purpose and research questions, followed by a discussion of problem statements.

Statements of Purpose

Many researchers articulate their goals as a statement of purpose, worded declaratively. The purpose statement establishes the study's general direction and captures its essence. It is usually easy to identify a purpose statement because the word *purpose* is explicitly stated: "The purpose of this study was . . ."—although sometimes the words *aim, goal, intent,* or *objective* are used instead, as in "The aim of this study was. . . ."

In a quantitative study, a statement of purpose identifies the key study variables and their possible interrelationships, as well as the population of interest.

Example of a statement of purpose from a quantitative study: "The primary purpose of this study was to determine the incidence of and associated risk for falls and fractures among adults 12 to 60 months after they underwent RYGB (Roux-en-Y gastric bypass) for morbid obesity" (Berarducci et al., 2009, p. 35).

This purpose statement identifies the population—individuals who have undergone RYGB surgery—and indicates two goals. The first is descriptive, that is, to describe the incidence of falls and fractures within the population. The second is to examine the effect of risk factors, such as use of analgesics, diuretics, and sedatives (the independent variables) on fall and fracture incidence (the dependent variables).

In qualitative studies, the statement of purpose indicates the key concept or phenomenon, and the group, community, or setting under study.

Example of a statement of purpose from a qualitative study: "The purpose of this study was to explore the characteristics of and the contexts related to sexual behaviors among institutionalized residents with dementia" (Tzeng et al., 2009, p. 991).

This statement indicates that the central phenomenon was the characteristics and contexts of sexual behavior, and that the group under study was institutionalized residents with dementia.

The statement of purpose communicates more than just the nature of the problem. Researchers' selection of verbs in a purpose statement suggests how they sought to solve the problem, or the state of knowledge on the topic. A study whose purpose is to *explore* or *describe* a phenomenon is likely an investigation of a little-researched topic, sometimes involving a qualitative approach such as a phenomenology or ethnography. A statement of purpose for a qualitative study—especially a grounded theory study—may also use verbs such as *understand, discover, develop,* or *generate.* Statements of purpose in qualitative studies may "encode" the tradition of inquiry, not only through the researcher's choice of verbs, but also through the use of "buzz words" associated with those traditions, as follows:

- *Grounded theory*: Processes, social structures, social interactions
- *Phenomenological studies*: experience, lived experience, meaning, essence
- *Ethnographic studies*: culture, roles, lifeways, cultural behavior

Quantitative researchers also suggest the nature of the inquiry through their selection of verbs. A statement indicating that the purpose of the study is to *test* or *evaluate* something (e.g., an intervention) suggests an experimental design, for example. A study whose

purpose is to *examine* or *explore* the relationship between two variables is more likely to involve a nonexperimental design. In some cases, the verb is ambiguous: a purpose statement indicating that the researcher's intent is to *compare* could be referring to a comparison of alternative treatments (using an experimental approach) or a comparison of two preexisting groups (using a nonexperimental approach). In any event, verbs such as *test*, *evaluate*, and *compare* suggest an existing knowledge base and quantifiable variables.

Note that the choice of verbs in a statement of purpose should connote objectivity. A statement of purpose indicating that the intent of the study was to *prove*, *demonstrate*, or *show* something suggests a bias.

TIP: In wording your statement of purpose, it may be useful to look at published research articles for models. Unfortunately, some reports fail to state unambiguously the study purpose, leaving readers to infer the purpose from such sources as the title of the report. In other reports, the purpose is clearly stated but may be difficult to find. Researchers most often state their purpose toward the end of the report's introduction.

Research Questions

Research questions are, in some cases, direct rewordings of statements of purpose, phrased interrogatively rather than declaratively, as in the following example:

- The purpose of this study is to assess the relationship between the dependency level of renal transplant recipients and their rate of recovery.
- What is the relationship between the dependency level of renal transplant recipients and their rate of recovery?

The question form has the advantage of simplicity and directness. Questions invite an answer and help to focus attention on the kinds of data that would have to be collected to provide that answer. Some research reports thus omit a statement of purpose and state only research questions. Other researchers use a set of research questions to clarify or lend greater specificity to a global purpose statement.

Research Questions in Quantitative Studies

In Chapter 2, we discussed the framing of clinical foreground questions to guide an EBP inquiry. Many of the EBP question templates in Table 2.1 could yield questions to guide a study as well, but *researchers* tend to conceptualize their questions in terms of their *variables*. Take, for example, the first question in Table 2.1, which states, "In (population), what is the effect of (intervention) on (outcome)? A researcher would likely think of the question in these terms: "In (population), what is the effect of (independent variable) on (dependent variable)? The advantage of thinking in terms of variables is that researchers must consciously decide how to operationalize their variables and how to guide an analysis strategy with their variables. Thus, we can say that in quantitative studies, research questions identify key study variables, the relationships among them, and the population under study. The variables are all measurable, quantifiable concepts.

Most research questions concern relationships among variables, and so many quantitative research questions could be articulated using a general question template: "In (population), what is the relationship between (independent variable or IV) and (dependent variable or DV)?" Examples of minor variations include the following:

- *Treatment, intervention*: In (population), what is the effect of (IV: intervention) on (DV)?
- *Prognosis*: In (population), does (IV: disease, condition) affect or increase the risk of (DV: adverse consequences)?
- *Causation, etiology*: In (population), does (IV: exposure, characteristic) cause or increase the risk of (DV: disease, health problem)?

There is one important distinction between the clinical foreground questions for an EBP-focused evidence search as described in Chapter 2 and a research question for an original study. As shown in Table 2.1, sometimes clinicians ask questions about explicit comparisons (e.g., they want to compare

intervention A to intervention B) and sometimes they do not (e.g., they want to learn the effects of intervention A, compared to any other intervention or to the absence of an intervention). In a research question, there must *always* be a designated comparison, because the independent variable must be operationally defined; this definition would articulate exactly what is being studied.

Another distinction between EBP and research questions is that research questions sometimes are more complex than clinical foreground questions for EBP. As an example, suppose that we began with an interest in nurses' use of humor with cancer patients, and the effects that humor has on these patients. One research question might be, "What is the effect of nurses' use of humor (versus absence of humor, the IV) on stress (the DV) in hospitalized cancer patients (the population)? But we might also be interested in whether the relationship between the IV and the DV is influenced by or *moderated* by a third variable. For example: Does nurses' use of humor have a different effect on stress in male versus female patients? In this example, gender is a **moderator variable**—a variable that affects the strength or direction of an association between the independent and dependent variable. Identifying moderators may be important in understanding *when* to expect a relationship between the IV and DV, and often has clinical relevance. Moderator (or *moderating*) variables can be characteristics of the population (e.g., male versus female patients) or of the circumstances (e.g., rural versus urban settings). Here are examples of question templates that involve a moderator variable (MV):

- *Treatment, intervention*: In (population), does the effect of (IV: intervention) on (DV) vary by (MV)?
- *Prognosis*: In (population), does the effect of (IV: disease, condition) on (DV) vary by (MV)?
- *Causation, etiology*: In (population), does (IV: exposure, characteristic) cause or increase risk of (DV) differentially by (MV)?

When a study purpose is to understand *causal pathways*, research questions may involve a **mediating variable**—a variable that intervenes between the IV and the DV and helps to explain why the relationship exists. In our example, we might ask the following: Does nurses' use of humor have a direct effect on the stress of hospitalized patients with cancer, or is the effect *mediated by* humor's effect on natural killer cell activity?

Some research questions are primarily descriptive. As examples, here are some descriptive questions that could be answered in a study on nurses' use of humor:

- What is the frequency with which nurses use humor as a complementary therapy with hospitalized cancer patients?
- What are the attitudes of hospitalized cancer patients to nurses' use of humor?
- What are the characteristics of nurses who use humor as a complementary therapy with hospitalized cancer patients?

Answers to such questions might, if addressed in a methodologically sound study, be useful in developing strategies for reducing stress in patients with cancer.

Example of a research question from a quantitative study: Robbins and colleagues (2009) studied gender differences in middle school children's attitudes toward physical activity. One of their key research questions was: *Do middle school boys and girls differ in their perceived benefits of and barriers to physical activity?*

TIP: The toolkit section of Chapter 4 of the accompanying *Resource Manual* includes a Word document that can be "filled in" to generate many types of research questions for both qualitative and quantitative studies.

Research Questions in Qualitative Studies

Research questions for qualitative studies state the phenomenon of interest and the group or population of interest. Researchers in the various qualitative traditions vary in their conceptualization of what types of questions are important. Grounded theory researchers are likely to ask *process* questions, phenomenologists tend to ask *meaning* questions, and ethnographers generally ask *descriptive*

questions about cultures. The terms associated with the various traditions, discussed previously in connection with purpose statements, are likely to be incorporated into the research questions.

Example of a research question from a phenomenological study: What is women's lived experience of fear of childbirth? (Nilsson & Lundgren, 2009).

Not all qualitative studies are rooted in a specific research tradition. Many researchers use qualitative methods to describe or explore phenomena without focusing on cultures, meaning, or social processes.

Example of a research question from a descriptive qualitative study: Horne and colleagues (2010) conducted a descriptive qualitative study that asked, *What do young older adults perceive to be the influence of primary healthcare professionals in encouraging exercise and physical activity?*

In qualitative studies, research questions may evolve over the course of the study. Researchers begin with a *focus* that defines the broad boundaries of the study, but the boundaries are not cast in stone. The boundaries "can be altered and, in the typical naturalistic inquiry, will be" (Lincoln & Guba, 1985, p. 228). The naturalist begins with a research question that provides a general starting point but does not prohibit discovery; qualitative researchers are sufficiently flexible that questions can be modified as new information makes it relevant to do so.

Problem Statements

Problem statements express the dilemma or troubling situation that needs investigation and that provides a rationale for a new inquiry. A problem statement identifies the nature of the problem that is being addressed and its context and significance. A problem statement is *not* merely a statement of the purpose of the study, it is a well-structured formulation of what it is that is problematic, what it is that "needs fixing," or what it is that is poorly understood. Problem statements, especially for quantitative studies, often have most of the following six components:

1. *Problem identification*: What is wrong with the current situation?
2. *Background*: What is the context of the problem that readers need to understand?
3. *Scope of the problem*: How big a problem is it, how many people are affected?
4. *Consequences of the problem*: What is the cost of not fixing the problem?
5. *Knowledge gaps*: What information about the problem is lacking?
6. *Proposed solution*: What is the basis for believing that the proposed study would contribute to the solution of the problem?

TIP: The toolkit section of Chapter 4 of the accompanying *Resource Manual* includes these questions in a Word document that can be "filled in" and reorganized as needed, as an aid to developing a problem statement.

Suppose our topic was humor as a complimentary therapy for reducing stress in hospitalized patients with cancer. Our research question is, "What is the effect of nurses' use of humor on stress and natural killer cell activity in hospitalized cancer patients?" Box 4.1 presents a rough draft of a problem statement for such a study. This problem statement is a reasonable first draft. The draft has several, but not all, of the six components.

Box 4.2 illustrates how the problem statement could be strengthened by adding information about scope (component 3), long-term consequences (component 4), and possible solutions (component 6). This second draft builds a more compelling argument for new research: millions of people are affected by cancer, and the disease has adverse consequences not only for those diagnosed and their families, but also for society. The revised problem statement also describes preliminary findings on which the new study might build.

As this example suggests, the problem statement is usually interwoven with supportive evidence from the research literature. In many research articles, it

BOX 4.1 Draft Problem Statement on Humor and Stress

A diagnosis of cancer is associated with high levels of stress. Sizeable numbers of patients who receive a cancer diagnosis describe feelings of uncertainty, fear, anger, and loss of control. Interpersonal relationships, psychological functioning, and role performance have all been found to suffer following cancer diagnosis and treatment.

A variety of alternative/complementary therapies have been developed in an effort to decrease the harmful effects of stress on psychological and physiological functioning, and resources devoted to these therapies (money and staff) have increased in recent years. However, many of these therapies have not been carefully evaluated to determine their efficacy, safety, or cost effectiveness. For example, the use of humor has been recommended as a therapeutic device to improve quality of life, decrease stress, and perhaps improve immune functioning, but the evidence to justify its popularity is scant.

is difficult to disentangle the problem statement from the literature review, unless there is a subsection specifically labeled "Literature Review."

Problem statements for a qualitative study similarly express the nature of the problem, its context, its scope, and information needed to address it, as in this example with bracketed citations:

Example of a problem statement from a qualitative study: "An unhealthy diet and lack of activity are two of the major risk factors responsible for increases in non-communicable diseases in modern societies. Problems such as cardiovascular and coronary heart disease, obesity, diabetes, and cancer account for more than half of deaths (60%) and nearly half (47%) of the burden of disease worldwide [1] . . . As prevention is a priority, the impact that children's activity levels and diet could have on their current and future health is of special concern [3] . . . Parents have a great influence on food [5] and activity [6,7] choices and behaviours of their offspring . . . This study used a qualitative design . . . to investigate how mothers and fathers contributed to food and activity choices and maintenance of a healthy lifestyle in children" (Lopez-Dicastillo et al., 2010).

BOX 4.2 Some Possible Improvements to Problem Statement on Humor and Stress

Each year, more than 1 million people are diagnosed with cancer, which remains one of the top causes of death among both men and women (citations). Numerous studies have documented that a diagnosis of cancer is associated with high levels of stress. Sizeable numbers of patients who receive a cancer diagnosis describe feelings of uncertainty, fear, anger, and loss of control (citations). Interpersonal relationships, psychological functioning, and role performance have all been found to suffer following cancer diagnosis and treatment (citations). These stressful outcomes can, in turn, adversely affect health, long-term prognosis, and medical costs among cancer survivors (citations).

A variety of alternative/complementary therapies have been developed in an effort to decrease the harmful effects of stress on psychological and physiological functioning, and resources devoted to these therapies (money and staff) have increased in recent years (citations). However, many of these therapies have not been carefully evaluated to determine their efficacy, safety, or cost effectiveness. For example, the use of humor has been recommended as a therapeutic device to improve quality of life, decrease stress, and perhaps improve immune functioning (citations), but the evidence to justify its popularity is scant. Preliminary findings from a recent small-scale endocrinology study with a healthy sample exposed to a humorous intervention (citation), however, holds promise for further inquiry with immunocompromised populations.

Qualitative studies that are embedded in a particular research tradition usually incorporate terms and concepts in their problem statements that foreshadow their tradition of inquiry (Creswell, 2006). For example, the problem statement in a grounded theory study might refer to the need to generate a theory relating to social processes. A problem statement for a phenomenological study might note the need to gain insight into people's experiences or the meanings they attribute to those experiences. And an ethnographer might indicate the need to understand how cultural forces affect people's behavior.

RESEARCH HYPOTHESES

A hypothesis is a prediction, almost always a prediction about the relationship between variables. In qualitative studies, researchers do not have an *a priori* hypothesis, in part because there is too little known to justify a prediction, and in part, because qualitative researchers want the inquiry to be guided by participants' viewpoints rather than by their own hunches. Thus, our discussion here focuses on hypotheses in quantitative research.

Function of Hypotheses in Quantitative Research

Research questions, as we have seen, are usually queries about relationships between variables. Hypotheses are predicted answers to these queries. For instance, the research question might ask: Does sexual abuse in childhood affect the development of irritable bowel syndrome in women? The researcher might predict the following: Women who were sexually abused in childhood have a higher incidence of irritable bowel syndrome than women who were not.

Hypotheses sometimes follow from a theoretical framework. Scientists reason from theories to hypotheses and test those hypotheses in the real world. The validity of a theory is evaluated through hypothesis testing. Take, as an example, the theory of reinforcement, which maintains that behavior that is positively reinforced (rewarded) tends to be learned or repeated. If the theory is valid, it should be possible to make predictions about human behavior. For example, the following hypothesis is deduced from reinforcement theory: Pediatric patients who are given a reward (e.g., a balloon or permission to watch television) when they cooperate during nursing procedures tend to be more cooperative during those procedures than nonrewarded peers. The theory gains support if the hypothesis is confirmed.

Not all hypotheses are derived from theory. Even in the absence of a theory, well-conceived hypotheses offer direction and suggest explanations. For example, suppose we hypothesized that the incidence of bradycardia in extremely low-birth-weight infants undergoing intubation and ventilation would be lower using the closed tracheal suction system (CTSS) than using the partially ventilated endotracheal suction method (PVETS). We could justify our speculation based on earlier studies or clinical observations, or both. *The development of predictions in and of itself forces researchers to think logically, to exercise critical judgment, and to tie together earlier research findings.*

Now, let us suppose the preceding hypothesis is not confirmed: We find that rates of bradycardia are similar for both the PVETS and CTSS methods. *The failure of data to support a prediction forces researchers to analyze theory or previous research critically, to carefully review the limitations of the study's methods, and to explore alternative explanations for the findings.* The use of hypotheses in quantitative studies tends to induce critical thinking and to facilitate understanding and interpretation of the data.

To illustrate further the utility of hypotheses, suppose we conducted the study guided only by the research question, Is there a relationship between suction method and rates of bradycardia? The investigator without a hypothesis is apparently prepared to accept any results. The problem is that it is almost always possible to explain something superficially after the fact, no matter what the findings are. Hypotheses guard against superficiality and minimize the risk that spurious results will be misconstrued.

Characteristics of Testable Hypotheses

Testable hypotheses state the expected relationship between the independent variable (the presumed cause or antecedent) and the dependent variable (the presumed effect or outcome) within a population.[1]

Example of a research hypothesis: Moore and co-researchers (2009) tested patency time in long-term indwelling urethral catheters among patients in three groups: those receiving standard care, a normal saline washout, or an acidic washout solution. The researchers hypothesized that time to first catheter change would be longest among patients who had the acidic washout solution.

In this example, the population is patients with long-term indwelling urethral catheters, the independent variable is method of managing blockages, and the dependent variable is the length of time elapsed until first catheter change. The hypothesis predicts that these two variables are related within the population—longer catheter life was expected for those receiving the acidic washout solution.

When researchers' hypotheses do not make a relational statement, the hypothesis is difficult to test. Take the following example: *Pregnant women who receive prenatal instruction regarding postpartum experiences are not likely to experience postpartum depression.* This statement expresses no anticipated relationship. There is only one variable (postpartum depression), and a relationship by definition requires at least two variables.

The problem is that without a prediction about an anticipated relationship, the hypothesis is difficult to test using standard procedures. In our example, how would we know whether the hypothesis was supported—what standard could be used to decide whether to accept or reject it? To illustrate this concretely, suppose we asked a group of mothers who had been given instruction on postpartum experiences the following question 1 month after delivery: On the whole, how depressed have you been since you gave birth? Would you say (1) extremely depressed, (2) moderately depressed, (3) a little depressed, or (4) not at all depressed?

Based on responses to this question, how could we compare the actual outcome with the predicted outcome? Would *all* the women have to say they were "not at all depressed?" Would the prediction be supported if 51% of the women said they were "not at all depressed" *or* "a little depressed?" It is difficult to test the accuracy of the prediction.

A test is simple, however, if we modify the prediction to the following: Pregnant women who receive prenatal instruction are less likely to experience postpartum depression than those with no prenatal instruction. Here, the dependent variable is the women's depression, and the independent variable is receipt versus nonreceipt of prenatal instruction. The relational aspect of the prediction is embodied in the phrase *less than.* If a hypothesis lacks a phrase such as *more than*, *less than*, *greater than*, *different from*, *related to*, *associated with*, or something similar, it is probably not amenable to testing in a quantitative study. To test this revised hypothesis, we could ask two groups of women with different prenatal instruction experiences to respond to the question on depression and then compare the groups' responses. The absolute degree of depression of either group would not be at issue.

Hypotheses should be based on justifiable rationales. Hypotheses often follow from previous research findings or are deduced from a theory. When a relatively new area is being investigated, the researcher may have to turn to logical reasoning or clinical experience to justify predictions.

The Derivation of Hypotheses

Many students ask, How do I go about developing hypotheses? Two basic processes—induction and deduction—are the intellectual machinery involved in deriving hypotheses.

An **inductive hypothesis** is a generalization inferred from observed relationships. Researchers observe certain patterns or associations among phenomena and then make predictions based on the observations. Related literature should be examined to learn what is known on a topic, but an

[1]It is possible to test hypotheses about the value of a single variable, but this happens rarely. See Chapter 17 for an example.

important source for inductive hypotheses is clinical experiences, combined with critical analysis. For example, a nurse might notice that presurgical patients who ask a lot of questions about pain or who express pain-related fears have a more difficult time than other patients in learning appropriate postoperative procedures. The nurse could formulate a testable hypothesis, such as: Patients who are stressed by fear of pain will have more difficulty in deep breathing and coughing after their surgery than patients who are not stressed. Qualitative studies are an important source of inspiration for inductive hypotheses.

Example of deriving an inductive hypothesis: In Beck and Watson's (2008) qualitative study on the impact of birth trauma on breastfeeding, one of their findings was that many mothers who had experienced birth trauma experienced intrusive, unwelcome flashbacks that caused them great distress. A hypothesis that can be derived from this qualitative finding might be as follows: Women who experience a traumatic childbirth have more flashbacks of their labor and delivery during breastfeeding than women who do not experience birth trauma.

Deduction is the other mechanism for deriving hypotheses. Theories of how phenomena interrelate cannot be tested directly but researchers can, through deductive reasoning, develop hypotheses based on theoretical principles. Inductive hypotheses begin with specific observations and move toward generalizations. **Deductive hypotheses** have theories as a starting point. Researchers ask: If this theory is valid, what are the implications for the variables of interest? Researchers deduce that if the general theory is true, then certain outcomes can be expected. Specific predictions derived from general principles must then be subjected to testing through data collection and analysis. If hypotheses are supported, then the theory is strengthened.

The advancement of nursing knowledge depends on both inductive and deductive hypotheses. Ideally, an iterative process is set in motion wherein observations are made (e.g., in a qualitative study), inductive hypotheses are formulated, systematic observations are made to test the hypotheses, theories are developed on the basis of the results, deductive hypotheses are formulated from the theory, new data are gathered, theories are modified, and so forth. Researchers need to be organizers of concepts (think inductively), logicians (think deductively), and critics and skeptics of resulting formulations, constantly demanding evidence.

Wording of Hypotheses

A good hypothesis is worded clearly and concisely, and in the present tense. Researchers make predictions about relationships that exist in the population, and not just about a relationship that will be revealed in a particular sample. There are various types of hypotheses.

Simple versus Complex Hypotheses

In this book, we define a **simple hypothesis** as a hypothesis that states an expected relationship between *one* independent and *one* dependent variable. A **complex hypothesis** is a prediction of a relationship between two or more independent variables and/or two or more dependent variables.

Simple hypotheses state a relationship between one independent variable, which we will call *X*, and one dependent variable, which we will call *Y*. *Y* is the predicted effect, outcome, or consequence of *X*, which is the presumed cause or antecedent. This relationship is shown graphically in Figure 4.1A. The circles represent variables *X* and *Y*, and the hatched area designates the strength of the relationship between them. If there were a one-to-one correspondence between *X* and *Y*, the two circles would overlap completely. If the variables were unrelated, the circles would not overlap at all. The previously cited study of catheter patency time in three catheter management groups (Moore et al., 2009) illustrates a simple hypothesis.

Most phenomena are affected by a multiplicity of factors. A person's weight, for example, is affected simultaneously by such factors as height, diet, bone structure, activity level, and metabolism. If *Y* in Figure 4.1A was weight, and *X* was a person's caloric intake, we would not be able to explain or understand individual variation in weight very well. For example, knowing that Nate

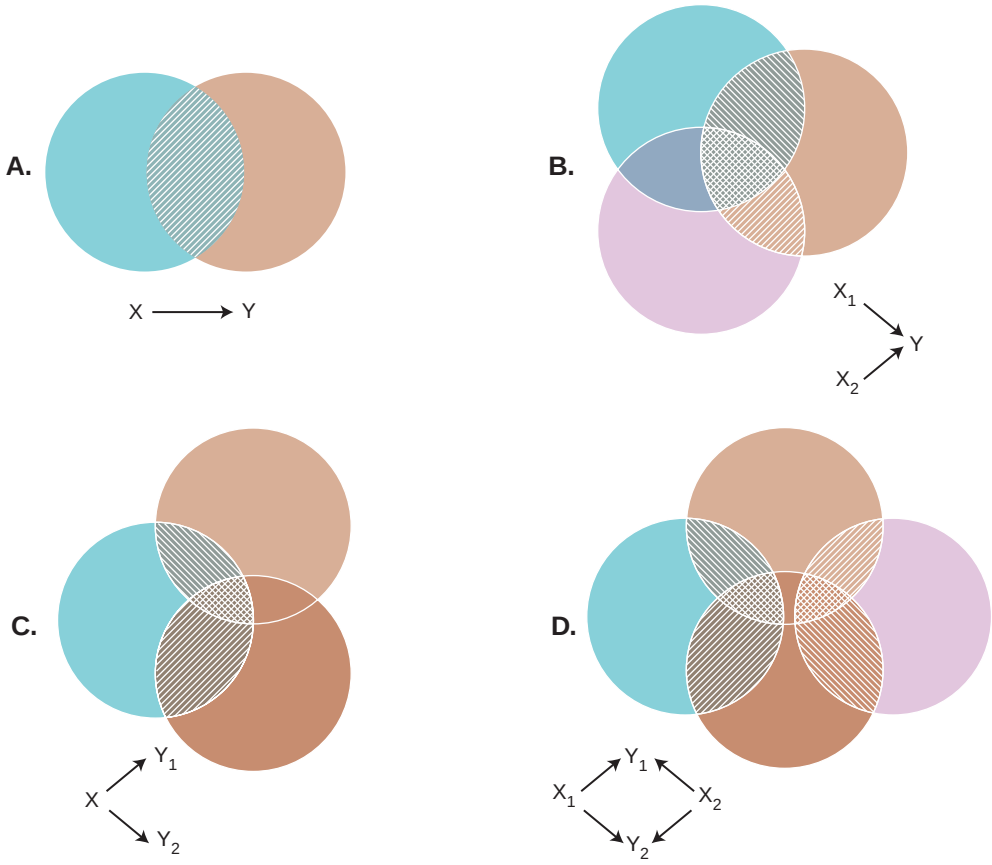


FIGURE 4.1 Schematic representation of various hypothetical relationships. (X = Independent variable; Y = Dependent variable.)

O'Hara's daily caloric intake averages 2,500 calories would not permit a good prediction of his weight. Knowledge of other factors, such as his height, would improve the accuracy with which his weight could be predicted.

Figure 4.1B presents a schematic representation of the effect of two independent variables (X_1 and X_2) on one dependent variable (Y). To pursue the preceding example, the hypothesis might be: Taller people (X_1) and people with higher caloric intake (X_2) weigh more (Y) than shorter people and those with lower caloric intake. As the figure shows, a larger proportion of the area of Y is hatched when there are two independent variables than when there is only one. This means that caloric intake *and* height do a better job in helping us explain variation in weight (Y) than caloric intake alone. Complex hypotheses have the advantage of allowing researchers to capture some of the complexity of the real world.

Just as a phenomenon can result from more than one independent variable, so a single independent variable can influence more than one phenomenon, as illustrated in Figure 4.1C. A number of studies have found, for example, that cigarette smoking (the independent variable, X), can lead to both lung cancer (Y_1) and coronary disorders (Y_2). Complex hypotheses are common in studies that

try to assess the impact of a nursing intervention on multiple outcomes.

Example of a complex hypothesis—multiple dependent variables: Lundberg and colleagues (2009) hypothesized that mental health patients who experienced stigmatizing rejection experiences [*X*] would, compared to those without such experiences, have lower self-esteem [Y_1], lower sense of empowerment [Y_2], and lower sense of coherence [Y_3].

A more complex type of hypothesis, which links two or more independent variables to two or more dependent variables, is shown in Figure 4.1D. An example might be a hypothesis that smoking *and* the consumption of alcohol during pregnancy might lead to lower birth weights *and* lower Apgar scores in infants.

Hypotheses are also complex if mediating or moderator variables are included in the prediction. For example, it might be hypothesized that the effect of caloric intake (*X*) on weight (*Y*) is moderated by gender (*Z*)—that is, the relationship between height and weight is different for men and women. Or, we might predict that the effect of ephedra (*X*) on weight (*Y*) is indirect, mediated by ephedra's effect on metabolism (*Z*).

Directional versus Nondirectional Hypotheses

Hypotheses can be stated in a number of ways, as in the following examples:

1. Older patients are more at risk of experiencing a fall than younger patients.
2. There is a relationship between the age of a patient and the risk of falling.
3. The older the patient, the greater the risk that he or she will fall.
4. Older patients differ from younger ones with respect to their risk of falling.
5. Younger patients tend to be less at risk of a fall than older patients.
6. The risk of falling increases with the age of the patient.

In each example, the hypothesis indicates the population (patients), the independent variable (patients' age), the dependent variable (a fall), and the anticipated relationship between them.

Hypotheses can be either directional or nondirectional. A **directional hypothesis** is one that specifies not only the existence but also the expected direction of the relationship between variables. In the six versions of the hypothesis, versions 1, 3, 5, and 6 are directional because there is an explicit prediction that older patients are at greater risk of falling than younger ones.

A **nondirectional hypothesis**, by contrast, does not state the direction of the relationship. Versions 2 and 4 in the example illustrate nondirectional hypotheses. These hypotheses state the prediction that a patient's age and risk of falling are related, but they do not stipulate whether the researcher thinks that *older* patients or *younger* ones are at greater risk.

Hypotheses derived from theory are almost always directional because theories provide a rationale for expecting variables to be related in a certain way. Existing studies also offer a basis for directional hypotheses. When there is no theory or related research, when findings of prior studies are contradictory, or when researchers' own experience leads to ambivalence, nondirectional hypotheses may be appropriate. Some people argue, in fact, that nondirectional hypotheses are preferable because they connote impartiality. Directional hypotheses, it is said, imply that researchers are intellectually committed to certain outcomes, and such a commitment might lead to bias. This argument fails to recognize that researchers typically *do* have hunches about outcomes, whether they state those expectations explicitly or not. We prefer directional hypotheses—when there is a reasonable basis for them—because they clarify the study's framework and demonstrate that researchers have thought critically about the phenomena under study. Directional hypotheses may also permit a more sensitive statistical test through the use of a *one-tailed test*—a rather fine point we discuss in Chapter 17.

Research versus Null Hypotheses

Hypotheses can be described as either research hypotheses or null hypotheses. **Research hypotheses**

(also called *substantive* or *scientific* hypotheses) are statements of expected relationships between variables. All hypotheses presented thus far are research hypotheses that indicate actual expectations.

Statistical inference uses a logic that may be confusing. This logic requires that hypotheses be expressed as an expected *absence* of a relationship. **Null hypotheses** (or *statistical hypotheses*) state that there is no relationship between the independent and dependent variables. The null form of the hypothesis used in our example might be: "Patients' age is unrelated to their risk of falling" or "Older patients are just as likely as younger patients to fall." The null hypothesis might be compared with the assumption of innocence of an accused criminal in English-based systems of justice: The variables are assumed to be "innocent" of any relationship until they can be shown "guilty" through appropriate statistical procedures. The null hypothesis represents the formal statement of this assumption of innocence.

TIP: Avoid stating hypotheses in null form in a proposal or a report, because this gives an amateurish impression. When statistical tests are performed, the underlying null hypothesis is assumed without being explicitly stated.

Hypothesis Testing

Researchers seek evidence through statistical analysis that their research hypotheses have a high probability of being correct. However, hypotheses are never *proved* through hypothesis testing; rather, they are *accepted* or *supported*. Findings are always tentative. Certainly, if the same results are replicated in numerous studies, then greater confidence can be placed in the conclusions. Hypotheses come to be increasingly supported with mounting evidence.

Let us look at why this is so. Suppose we hypothesized that height and weight are related. We predict that, on average, tall people weigh more than short people. We then obtain height and weight measurements from a sample and analyze the data. Now, suppose we happened by chance to get a sample that consisted of short, heavy people, and tall, thin people. Our results might indicate that there is no relationship between height and weight. Would we be justified in stating that this study *proved* that height and weight are unrelated?

As another example, suppose we hypothesized that tall nurses are more effective than short ones. In reality, we would expect no relationship between height and a nurse's job performance. Now, suppose that, by chance again, we drew a sample in which tall nurses received better job evaluations than short ones. Could we conclude that height is related to a nurse's performance? These two examples illustrate the difficulty of using observations from a sample to generalize to a population. Other issues, such as the accuracy of the measures and the effects of uncontrolled variables prevent researchers from concluding with finality that hypotheses are proved.

TIP: If a researcher uses any statistical tests (as is true in most quantitative studies), it means that there are underlying hypotheses—regardless of whether the researcher explicitly stated them—because statistical tests are designed to test hypotheses. In planning a quantitative study of your own, do not be afraid to make predictions, that is, to state hypotheses.

CRITIQUING RESEARCH PROBLEMS, RESEARCH QUESTIONS, AND HYPOTHESES

In critiquing research articles, you need to evaluate whether researchers have adequately communicated their problem. The delineation of the problem, purpose statement, research questions, and hypotheses sets the stage for the description of what was done and what was learned. Ideally, you should not have to dig too deeply to decipher the research problem or to discover the questions.

A critique of the research problem is multidimensional. Substantively, you need to consider whether the problem is significant and has the potential to produce evidence to improve nursing practice. Studies that build in a meaningful way on existing knowledge are well-poised to contribute to

BOX 4.3 Guidelines for Critiquing Research Problems, Research Questions, and Hypotheses

1. What is the research problem? Is the problem statement easy to locate and is it clearly stated? Does the problem statement build a cogent and persuasive argument for the new study?
2. Does the problem have significance for nursing? How might the research contribute to nursing practice, administration, education, or policy?
3. Is there a good fit between the research problem and the paradigm within which the research was conducted? Is there a good fit between the problem and the qualitative research tradition (if applicable)?
4. Does the report formally present a statement of purpose, research question, and/or hypotheses? Is this information communicated clearly and concisely, and is it placed in a logical and useful location?
5. Are purpose statements or questions worded appropriately? For example, are key concepts/variables identified and is the population of interest specified? Are verbs used appropriately to suggest the nature of the inquiry and/or the research tradition?
6. If there are no formal hypotheses, is their absence justified? Are statistical tests used in analyzing the data despite the absence of stated hypotheses?
7. Do hypotheses (if any) flow from a theory or previous research? Is there a justifiable basis for the predictions?
8. Are hypotheses (if any) properly worded—do they state a predicted relationship between two or more variables? Are they directional or nondirectional, and is there a rationale for how they were stated? Are they presented as research or as null hypotheses?

evidence-based nursing practice. Researchers who develop a systematic *program of research*, building on their own earlier findings, are especially likely to make important contributions (Conn, 2004). For example, Beck's series of studies relating to postpartum depression have influenced women's healthcare worldwide. Also, research problems stemming from established research priorities (Chapter 1) have a high likelihood of yielding important new evidence for nurses because they reflect expert opinion about areas of needed research.

Another dimension in critiquing the research problem is methodologic—in particular, whether the research problem is compatible with the chosen research paradigm and its associated methods. You should also evaluate whether the statement of purpose or research questions have been properly worded and lend themselves to empirical inquiry.

In a quantitative study, if the research article does not contain explicit hypotheses, you need to consider whether their absence is justified. If there are hypotheses, you should evaluate whether they are logically connected to the problem and are consistent with existing evidence or relevant theory. The wording of hypotheses should also be assessed. To be testable, the hypothesis should contain a prediction about the relationship between two or more measurable variables. Specific guidelines for critiquing research problems, research questions, and hypotheses are presented in Box 4.3.

RESEARCH EXAMPLES

This section describes how the research problem and research questions were communicated in two nursing studies, one quantitative and one qualitative.

Research Example of a Quantitative Study

Study: The relationship among self-esteem, stress, coping, eating behavior, and depressive mood in adolescents (Martyn-Nemeth et al., 2009).

Problem Statement: "The prevalence of adolescent overweight has increased from 5% to 17% over the past 30 years in the United States . . . There are serious long-term health consequences for adolescents who are overweight . . . In addition, all overweight adolescents are at increased risk for depressive mood and clinical depression. Overweight adolescents tend to remain overweight as adults, with an increased risk of diabetes, cardiovascular disease, and cancer . . . The overall estimated economic burden of obesity in the nation for the year 2002 was 93 billion dollars . . . Self-esteem is associated with overeating and weight gain in adolescents, and stress-induced eating and inadequate coping skills have been related to overeating and obesity in adults . . . Important questions remain about the relationship of self-esteem, stress, social support, and coping to eating patterns in racially/ethnically diverse male and female adolescents" (p. 98).

Statement of Purpose: The purpose of this study "was to examine relationships among self-esteem, stress, social support, and coping, and to test a model of their effects on eating behavior and depressive mood in a sample of high school students" (p. 96).

Research Questions: The authors posed three research questions about relationships among the study variables (e.g., "Does the use of food as a coping mechanism relate to being overweight?" p. 99) One question focused on a mediating variable: "Does coping mediate the relationship of low self-esteem, increased stress, and decreased social support with the outcomes of unhealthy eating behavior and depressive mood" (p. 99).

Hypotheses: It was hypothesized that adolescents with low self-esteem, increased stress, and decreased social support would predominantly use avoidance mechanisms of coping, which would in turn mediate the negative outcomes of unhealthy eating and depressive mood.

Study Methods: The study was conducted with a multiracial sample of 102 students from two public high schools in Midwestern United States. Data were collected through self-administered questionnaires.

Key Findings: The results indicated that low self-esteem and stress were related to avoidant coping and depressive mood. Also, low self-esteem and avoidant coping were related to unhealthy eating, thus offering partial support for the researchers' hypotheses.

Research Example of a Qualitative Study

Study: Sustaining self: The lived experience of transition to long-term ventilation (Briscoe & Woodgate, 2010).

Problem Statement: "Chronic respiratory failure (CRF) occurs as a result of irreversible and/or progressive deterioration in ventilation and gas exchange, and is a common end point of a number of conditions that affect the lung, chest wall, and/or neurologic system . . . The only treatment for CRF is mechanical ventilation (MV), which can be delivered invasively via a tracheotomy tube, or noninvasively via a tightly sealed nasal or face mask, mouthpiece, or negative-chest-pressure device . . . A consensus of measuring incidence of CRF and prevalence of ventilator utilization is reflected in the literature . . . Care for individuals requiring long-term mechanical ventilation (LTMV) is evolving, and there is growing impetus to comprehensively address operational, financial, ethical, and client-centered concerns . . . Gaining a comprehensive understanding of both the burdens and benefits of ventilator treatment is vital for health professionals, ventilator users, and families . . . Especially lacking is an understanding of their transition, or journey, from spontaneous breathing to the stable reliance on LTMV" (pp. 57–58) (Citations were omitted to streamline the presentation).

Statement of Purpose: "The purpose of this phenomenological study was to acquire a detailed description of the experience of transition to LTMV from individuals requiring ventilation" (p. 58). (No specific research questions were articulated in this article).

Method: Study participants were 11 ventilated individuals recruited from two respiratory care facilities in western Canada. All participants were interviewed on one or more occasions, and all interviews were audiorecorded. Participants shared pictures and other memorabilia, which assisted them in telling their stories of transition to LTMV. Conversational questions were posed, such as "Can you please tell me about the time when the ventilator was first introduced to you? Analysis began with the first interview and continued with ongoing interviews over a 4-month period.

Key Findings: The transition journey was found to be a time of psychological, physical, and spiritual challenge. "Sustaining self" was identified as the essence of ventilator users' transition experience.

SUMMARY POINTS

- A **research problem** is a perplexing or enigmatic situation that a researcher wants to address through disciplined inquiry. Researchers usually

identify a broad **topic,** narrow the problem scope, and identify questions consistent with a paradigm of choice.

- Common sources of ideas for nursing research problems are clinical experience, relevant literature, quality improvement initiatives, social issues, theory, and external suggestions.
- Key criteria in assessing a research problem are that the problem should be clinically significant; researchable; feasible; and of personal interest.
- Feasibility involves the issues of time, cooperation of participants and other people, availability of facilities and equipment, researcher experience, and ethical considerations.
- Researchers communicate their aims as problem statements, statements of purpose, research questions, or hypotheses.
- A **statement of purpose**, which summarizes the overall study goal, identifies key concepts (variables) and the population. Purpose statements often communicate, through the use of verbs and other key terms, the underlying research tradition of qualitative studies, or whether study is experimental or nonexperimental in quantitative ones.
- A **research question** is the specific query researchers want to answer in addressing the research problem. In quantitative studies, research questions usually concern the existence, nature, strength, and direction of relationships.
- Some research questions are about **moderator variables** that affect the strength or direction of a relationship between the independent and dependent variables; others are about **mediating variables** that intervene between the independent and dependent variable and help to explain why the relationship exists.
- **Problem statements,** which articulate the nature, context, and significance of a problem, include several components: problem identification; the background, scope, and consequences of the problem; knowledge gaps; and possible solutions to the problem.
- In quantitative studies, a **hypothesis** is a statement of predicted relationships between two or more variables.
- **Simple hypotheses** express a predicted relationship between one independent variable and one dependent variable, whereas **complex hypotheses** state an anticipated relationship between two or more independent variables and two or more dependent variables (or state predictions about mediating or moderator variables).
- **Directional hypotheses** predict the direction of a relationship; **nondirectional hypotheses** predict the existence of relationships, not their direction.
- **Research hypotheses** predict the existence of relationships; **null hypotheses,** which express the absence of a relationship, are the hypotheses subjected to statistical testing.
- Hypotheses are never proved or disproved in an ultimate sense—they are accepted or rejected, supported or not supported by the data.

STUDY ACTIVITIES

Chapter 4 of the *Resource Manual for Nursing Research: Generating and Assessing Evidence for Nursing Practice, 9th ed.*, offers study suggestions for reinforcing concepts presented in this chapter. In addition, the following questions can be addressed in classroom or online discussions:

1. Think of a frustrating experience you have had as a nursing student or as a practicing nurse. Identify the problem area. Ask yourself a series of questions until you have one that you think is researchable. Evaluate the problem in terms of the evaluation criteria discussed in this chapter.
2. To the extent possible, use the critiquing questions in Box 4.3 to appraise the research problems for the two studies used as research examples at the end of this chapter.

STUDIES CITED IN CHAPTER 4

Beck, C. T., & Watson, S. (2008). The impact of birth trauma on breastfeeding: A tale of two pathways. *Nursing Research, 57*, 228–236.

Berarducci, A., Haines, K., & Murr, M. (2009). Incidence of bone loss, falls, and fractures after Roux-en-Y gastric bybass for morbid obesity. *Applied Nursing Research, 22*, 35–41.

Briscoe, W., & Woodgate, R. (2010). Sustaining self: The lived experience of transition to long-term ventilation. *Qualitative Health Research, 20*, 57–67.

Horne, M., Skelton, D., Speed, S., & Todd, C. (2010). The influence of primary health care professionals in encouraging exercise and physical activity uptake among white and South Asian older adults. *Patient Education & Counseling, 78*, 97–103.

Lopez-Dicastillo, O., Grande, G., & Callery, P. (2010). Parents' contrasting views on diet versus activity of children. *Patient Education and Counseling, 78*, 117–123.

Lundberg, B., Hansson, L., Wentz, E., & Bjorkman, T. (2009). Are stigmatizing experiences among persons with mental illness related to perceptions of self-esteem, empowerment, and sense of coherence? *Journal of Psychiatric & Mental Health Nursing, 16*, 516–522.

Martyn-Nemeth, P., Penckofer, S., Gulanick, M., Velsor-Friedrich, B., & Bryant, F. (2009). The relationship among self-esteem, stress, coping, eating behavior, and depressive mood in adolescents. *Research in Nursing & Health, 32*, 96–109.

Moore, K., Hunter, H., McGinnis, R., Bascu, C., Fader, M., Gray, M., Getliffe, K., Chobanuk, J., Puttagunta, L., & Voaklander, D. C. (2009). Do catheter washouts extend patency time in long-term indwelling urethral catheters? *Journal of Wound, Ostomy, and Continence Nursing, 36*, 82–90.

Nilsson, C., & Lundgren, I. (2009). Women's lived experience of fear of childbirth. *Midwifery, 25*, 1–9.

Robbins, L., Sikorski, A., Hamel, L., Wu, T., & Wilbur, J. (2009). Gender comparisons of perceived benefits of and barriers to physical activity in middle school youth. *Research in Nursing & Health, 32*, 163–176.

Tzeng, Y. L., Lin, L., Shyr, Y., & Wen, J. (2009). Sexual behavior of institutionalized residents with dementia. *Journal of Clinical Nursing, 18*, 991–1001.

Methodologic and nonresearch references cited in this chapter can be found in a separate section at the end of the book.

5 Literature Reviews: Finding and Critiquing Evidence

Researchers typically conduct research within the context of existing knowledge by undertaking a thorough **literature review**. This chapter describes activities associated with literature reviews, including locating and critiquing studies. Many of these activities overlap with early steps in an EBP project, as described in Chapter 2.

GETTING STARTED ON A LITERATURE REVIEW

Before discussing the steps involved in doing a research-based literature review, we briefly discuss some general issues. The first concerns the viewpoint of qualitative researchers.

Literature Reviews in Qualitative Research Traditions

As noted in Chapter 3, qualitative researchers have varying opinions about reviewing the literature before doing a new study. Some of the differences reflect viewpoints associated with qualitative research traditions.

Grounded theory researchers often collect their data before reviewing the literature. The grounded theory takes shape as data are analyzed. Researchers then turn to the literature when the theory is sufficiently developed, seeking to relate prior findings to the theory. Glaser (1978) warned that, "It's hard enough to generate one's own ideas without the 'rich' detailment provided by literature in the same field" (p. 31). Thus, grounded theory researchers may defer a literature review, but then consider how previous research fits with or extends the emerging theory. McGhee and colleagues (2007), however, have noted how researchers can use reflexivity (a concept discussed at length later in this book) to prevent prior knowledge from distorting grounded theory analysis.

Phenomenologists often undertake a search for relevant materials at the outset of a study. In reviewing the literature, phenomenological researchers look for experiential descriptions of the phenomenon being studied (Munhall, 2012). The purpose is to expand the researcher's understanding of the phenomenon from multiple perspectives, and this may include an examination of artistic sources in which the phenomenon is described (e.g., in novels or poetry).

Even though "ethnography starts with a conscious attitude of almost complete ignorance" (Spradley, 1979, p. 4), literature that led to the choice of the cultural problem to be studied is often reviewed before data collection. A second, more thorough literature review is often done during data analysis and interpretation so that findings can be compared with previous findings.

Regardless of tradition, if funding is sought for a qualitative project, an upfront literature review is usually necessary. Reviewers need to understand the context for the proposed study, and must be persuaded that it should be funded.

Purposes and Scope of Research Literature Reviews

Written literature reviews are undertaken for many different purposes. The length of the product depends on its purpose. Regardless of length, a good review requires thorough familiarity with available evidence. As Garrard (2006) advised, you must strive to *own* the literature on a topic to be confident of preparing a state-of-the-art review. The major types of written research review include the following:

- *A review in a research report*. Literature reviews in the introduction to a report provide readers with an overview of existing evidence, and contribute to the argument for the new study. These reviews are usually only 2 to 4 double-spaced pages, and so, only key studies can be cited. The emphasis is on summarizing and evaluating an overall body of evidence.
- *A review in a proposal.* A literature review in a proposal provides context, confirms the need for new research, and demonstrates the writer's "ownership" of the literature. The length of such reviews is established in proposal guidelines, but is often just a few pages. This means that the review must reflect expertise on the topic in a very succinct fashion.
- *A review in a thesis or dissertation*. Dissertations in the traditional format (see Chapter 28) often include a thorough, critical literature review. An entire chapter may be devoted to the review, and such chapters are often 15 to 25 pages long. These reviews typically include an evaluation of the overall body of literature as well as critiques of key individual studies.
- *Free-standing literature reviews*. Nurses also prepare reviews that critically appraise and summarize a body of research, sometimes for a course or for an EBP project. Researchers who are experts in a field also may do systematic reviews that are published in journals (Chapter 27). Free-standing reviews are usually 15 to 25 pages long.

This chapter focuses on the preparation of a review as a component of an original study, but most activities are similar for other types of review. By doing a thorough review, researchers can determine how best to make a contribution to existing evidence—for example, whether there are gaps or inconsistencies in a body of research, or whether a replication with a new population is the right next step. A literature review also plays a role at the end of the study when researchers try to make sense of their findings.

Types of Information for a Research Review

Written materials vary in their quality and the kind of information they contain. In performing a literature review, you will have to decide what to read and what to include in a written review. We offer some suggestions that may help in making such decisions.

The most important type of information for a research review is findings from prior studies. You should rely mostly on **primary source** research reports, which are descriptions of studies written by the researchers who conducted them.

Secondary source research documents are descriptions of studies prepared by someone other than the original researcher. Literature reviews, for example, are secondary sources. If reviews are recent, they are a good place to start because they provide an overview of the topic and a valuable bibliography. Secondary sources are not substitutes for primary sources because they typically fail to provide much detail about studies, and are seldom completely objective.

TIP: For an EBP project, a recent, high-quality review may be sufficient to provide needed information about existing evidence, although it is wise to search for recent studies not covered by the review.

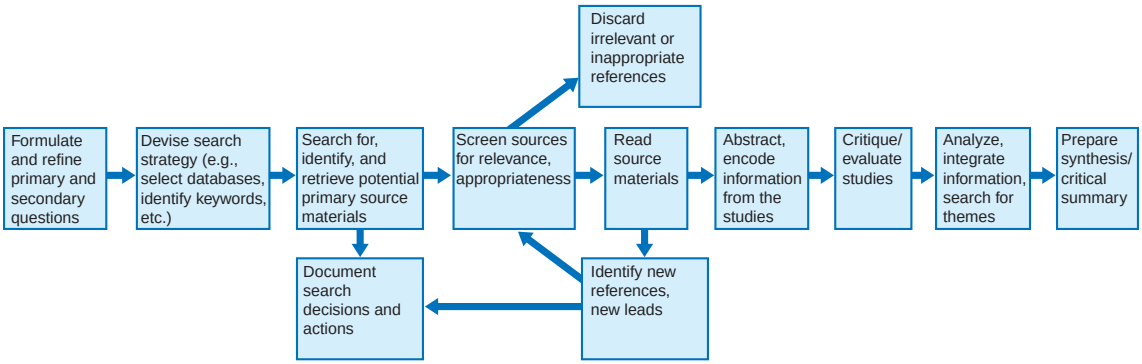


FIGURE 5.1 Flow of tasks in a literature review.

Examples of primary and secondary sources:

- Primary source, an original study of palliative patients and family caregivers regarding preferences for location of death: Stajduhar, K., Allan, D., Cohen, S., & Heyland, D. (2008). Preferences for location of death of seriously ill hospitalized patients. *Palliative Medicine,* 22, 85–88.
- Secondary source, a review of factors affecting place of end-of-life care for patients with cancer: Murray, M., Fiset, V., Young, S., & Kryworuchko, J. (2009). Where the dying live: Review of determinants of place of end-of-life cancer care. *Oncology Nursing Forum,* 36, 69–77.

In addition to research reports, your search may yield nonresearch references, such as case reports, anecdotes, or clinical descriptions. Nonresearch materials may broaden understanding of a problem, demonstrate a need for research, or describe aspects of clinical practice. These writings may help in formulating research ideas, but they usually have limited utility in written research reviews because they do not address the central question: What is the current state of *evidence* on this research problem?

Major Steps and Strategies in Doing a Literature Review

Conducting a literature review is a little like doing a full study, in the sense that reviewers start with a question, formulate and implement a plan for gathering information, and then analyze and interpret information. The "findings" must then be summarized in a written product.

Figure 5.1 outlines the literature review process. As the figure shows, there are several potential feedback loops, with opportunities to retrace earlier steps in search of more information. This chapter discusses each step, but some steps are elaborated in Chapter 27 in our discussion of systematic reviews.

Conducting a high-quality literature review is more than a mechanical exercise—it is an art and a science. Several qualities characterize a high-quality review. First, the review must be comprehensive, thorough, and up-to-date. To "own" the literature (Garrard, 2006), you must be determined to become an expert on your topic, which means that you need to be creative and diligent in hunting down leads for possible sources of information.

TIP: Locating all relevant information on a research question is a bit like being a detective. The literature retrieval tools we discuss in this chapter are a tremendous aid, but there inevitably needs to be some digging for the clues to evidence on a topic. Be prepared for sleuthing!

Second, a high-quality review is systematic. Decision rules should be clear, and criteria for including or excluding a study need to be explicit. This is because a third characteristic of a good review is that it is reproducible, which means that

another diligent reviewer would be able to apply the same decision rules and criteria and come to similar conclusions about the evidence.

Another desirable attribute of a literature review is the absence of bias. This is more easily achieved when systematic rules for evaluating information are followed—although reviewers cannot totally elude personal opinions. For this reason, systematic reviews are often conducted by teams of researchers who can evaluate each other's conclusions. Finally, reviewers should strive for a review that is insightful and that is more than "the sum of its parts." Reviewers have an opportunity to contribute to knowledge through an astute and incisive synthesis of the evidence.

We recommend thinking of doing a literature review as similar to doing a qualitative study. This means having a flexible approach to "data collection" and thinking creatively about ideas for new sources of information. It means pursuing leads until "saturation" is achieved—that is, until your search strategies yield redundant information about studies to include. And it also means that the analysis of your "data" will typically involve a search for important themes.

Primary and Secondary Questions for a Review

For free-standing literature reviews and EBP projects, the reviewer may seek to summarize research evidence about a single focused question, such as those described in Chapter 2 (see Table 2.1 for question templates). For those who are undertaking a literature review as part of a new study, the *primary question* for the literature review is the same as the actual research question for the new study. The researcher wants to know: What is the current state of knowledge on the question that I will be addressing in *my* study?

If you are doing a review for a new study, you inevitably will need to search for existing evidence on several *secondary questions* as well because you will need to develop an argument (a rationale) for the new study in the problem statement. An example (which we will use throughout this chapter) will clarify this point.

Suppose that we were conducting a study to address the following question: What characteristics of nurses are associated with effective pain management for hospitalized children? In other words, our primary question is whether there are characteristics of nurses that are associated with appropriate responses to children's pain. Such a question would arise within the context of a perceived problem, such as a concern that nurses' treatment of children's pain is not always optimal. A basic statement of the problem might be as follows:

• • •

> Many children are hospitalized annually and many hospitalized children experience high levels of pain. There are long-lasting harmful effects to the nervous system when severe or persistent pain in children is untreated. Although effective analgesic and nonpharmacologic methods of controlling children's pain exist, and although there are reliable methods of assessing children's pain, nurses do not always manage children's pain effectively. *What characteristics distinguish nurses who are effective and those who are not?*

• • •

This rudimentary problem statement suggests a number of *secondary questions* for which evidence from the literature will need to be located and evaluated. Examples of such secondary questions include the following:

- How many children are hospitalized annually?
- What types and levels of pain do hospitalized children experience?
- What are the consequences of untreated pain in children?
- How can pain in hospitalized children be reliably assessed and effectively treated?
- How adequately do nurses manage pain in hospitalized pediatric patients?

Thus, conducting a literature review tends to be a multipronged endeavor when it is done as part of a new study. While most of the "detective work" in searching the literature that we describe in this chapter applies principally to the primary question, it is important to keep in mind other questions for which information from the research literature needs to be retrieved.

LOCATING RELEVANT LITERATURE FOR A RESEARCH REVIEW

As shown in Figure 5.1, an early step in a literature review is devising a strategy to locate relevant studies. The ability to locate research documents on a topic is an important skill that requires adaptability. Rapid technological changes have made manual methods of finding information obsolete, and sophisticated methods of searching the literature are being introduced continuously. We urge you to consult with librarians, colleagues, or faculty for suggestions.

Formulating a Search Strategy

There are many ways to search for research evidence, and it is wise to begin a search with some strategies in mind. Cooper (2010) has identified several approaches, one of which we describe in some detail in this chapter: searching for references in bibliographic databases. Another approach, called the *ancestry approach*, involves using citations from relevant studies to track down earlier research on the same topic (the "ancestors"). A third method, the *descendancy approach*, is to find a pivotal early study and to search forward in citation indexes to find more recent studies ("descendants") that cited the key study. Other strategies exist for tracking down what is called the *grey literature*, which refers to studies with more limited distribution, such as conference papers, unpublished reports, and so on. We describe these strategies in Chapter 27 on systematic reviews. If your intent is to "own" the literature, then you will likely want to adopt all of these strategies, but in many cases, the first two or three might suffice.

TIP: You may be tempted to begin a literature search through an Internet search engine, such as Yahoo, Google, or Google Scholar. Such a search is likely to yield a lot of "hits" on your topic, but is not likely to give you full bibliographic information on *research* literature on your topic—and you might become frustrated with searching through vast numbers of website links.

Search plans also involve decisions about delimiting the search. These decisions need to be explicit to ensure reproducibility. If you are not multilingual, you may need to constrain your search to studies written in your own language. You may also want to limit your search to studies conducted within a certain time frame (e.g., within the past 15 years). You may want to exclude studies with certain types of participants. For instance, in our example of a literature search about nurses' characteristics and treatment of children's pain, we might want to exclude studies in which the children were neonates. Finally, you may choose to limit your search based on how your key variables are defined. For instance, in our example, you may (or may not) wish to exclude studies in which the focus was on nurses' *attitudes* toward children's pain.

TIP: Constraining your search might help you to avoid irrelevant material, but be cautious about putting too many restrictions on your search, especially initially. You can always make decisions to exclude studies at a later point, provided you have clear criteria and a rationale. Be sure not to limit your search to very recent studies or to studies exclusively in the nursing literature.

Searching Bibliographic Databases

Reviewers typically begin by searching bibliographic databases that can be accessed by computer. The databases contain entries for thousands of journal articles, each of which has been coded to facilitate retrieval. For example, articles may be coded for language used (e.g., English), subject matter (e.g., pain), type of journal (e.g., nursing), and so on. Several commercial vendors (e.g., Aries Knowledge Finder, Ovid, EBSCOhost, ProQuest) offer software for retrieving information from these databases. Most programs are user-friendly, offering menu-driven systems with on-screen support so that retrieval can proceed with minimal instruction. Some providers offer discount rates for students and trial services that allow you to test them before subscribing. In most cases, however, your university or hospital library has a subscription.

Getting Started with a Bibliographic Database

Before searching an electronic database, you should become familiar with the features of the software you are using to access the database. The software gives you options for limiting your search, for combining the results of two searches, for saving your search, and so on. Most programs have tutorials that can improve the efficiency and effectiveness of your search. In many cases, a "Help" button will provide you with a lot of information.

You will also need to learn how to get from "point A" (the constructs in which you are interested) to "point B" (the way that the program stores and organizes information about the constructs). Most software you are likely to use has mapping capabilities. *Mapping* is a feature that allows you to search for topics using your own **keywords**, rather than needing to enter a term that is exactly the same as a **subject heading** (subject codes) in the database. The software translates ("maps") the keywords you enter into the most plausible subject heading. In addition to mapping your term onto a database-specific subject heading, most programs will also search in the *text fields* of records (usually the title and abstract) for the keyword entered.

TIP: The keywords you begin with are usually your key independent or dependent variables, and perhaps your population. If you have used the question templates in Table 2.1 or in the Toolkit for Chapter 4, the words you entered in the blanks would be keywords.

Even when there are mapping capabilities, you should learn the relevant subject headings of the database you are using because keyword searches and subject heading searches yield overlapping but nonidentical results. Subject headings for databases can be located in the database's thesaurus or other reference tools.

TIP: To identify all major research reports on a topic, you need to be flexible and to think broadly about the keywords that could be related to your topic. For example, if you are interested in anorexia nervosa, you might look under *anorexia, eating disorder,* and *weight loss,* and perhaps under *appetite, eating behavior, food habits, bulimia,* and *body weight change.*

General Database Search Features

Some features of an electronic search are similar across databases. One feature is that you usually can use **Boolean operators** to expand or delimit a search. Three widely used Boolean operators are AND, OR, and NOT (usually in all caps). The operator *AND* delimits a search. If we searched for *pain AND children,* the software would retrieve only records that have both terms. The operator *OR* expands the search: *pain OR children* could be used in a search to retrieve records with either term. Finally, *NOT* narrows a search: *pain NOT children* would retrieve all records with pain that did not include the term children.

Wildcard and truncation symbols are other useful tools for searching databases. These symbols vary from one database to another, but their function is to expand the search. A **truncation symbol** (often an asterisk, *) expands a search term to include all forms of a root word. For example, a search for *child** would instruct the computer to search for any word that begins with "child" such as children, childhood, or childrearing. **Wildcard symbols** (often a question mark or asterisk) inserted into the middle of a search term permits a search for alternative spellings. For example, a search for *behavio?r* would retrieve records with either *behavior* or *behaviour.* Also, a search for *wom?n* would retrieve records with either *woman* or *women.* For each database, it is important to learn what these special symbols are and how they work. For example, many databases require at least three letters at the beginning of a search term before a wildcard or truncation symbol can be used (e.g., ca* would not be allowed). Moreover, not every database (including PubMed) allows wildcard codes in the middle of a search term.

Another important thing to know is that use of special symbols usually turns off a software's mapping feature. For example, a search for *child** would retrieve records in which any form of "child" appeared in text fields, but it would not map any of these concepts onto the database's subject headings (e.g., pediatric).

Sometimes it is important to keep words together in a search, as in a search for records with *blood*

pressure. Some bibliometric software would treat this as *blood AND pressure*, and would search for records with both terms somewhere in text fields, even if they are not contiguous. Quotation marks often can be used to ensure that the words are searched only in combination, as in "*blood pressure.*"

Key Electronic Databases for Nurse Researchers

Two especially useful electronic databases for nurse researchers are CINAHL (**C**umulative **I**ndex to **N**ursing and **A**llied **H**ealth **L**iterature) and MEDLINE (**Med**ical Literature On-**Line**), which we discuss in the next sections. Other potentially useful bibliographic databases for nurses include:

- British Nursing Index
- Cochrane Database of Systematic Reviews
- Dissertation Abstracts online
- EMBASE (the **E**xcerpta **M**edica data**base**)
- HaPI (**H**ealth **a**nd **P**sychosocial **I**nstruments database)
- Health Source: Nursing/Academic Edition
- ISI Web of Knowledge
- Nursing and Allied Health Source (ProQuest)
- PsycINFO (**Psyc**hology **Info**rmation)
- Scopus

Note that a search strategy that works well in one database does not always produce good results in another. Thus, it is important to explore strategies in each database and to understand how each database is structured—for example, what subject headings are used and how they are organized in a hierarchy. Each database and software program also has certain peculiarities. For example, using PubMed (to be discussed later) to search the MEDLINE database, you might restrict your search to nursing journals. However, if you did this you would be excluding studies in several journals in which nurses often publish, such as *Birth* and *Qualitative Health Research* because these journals are not coded for the nursing subset of PubMed.

TIP: In the next two sections, we provide specific information about using CINAHL and MEDLINE via PubMed. Note, however, that databases and the software through which they are accessed change from time to time, and our instructions may not be precisely accurate. For example, a redesigned interface was implemented in PubMed in late 2009 and was later revised in February 2010, requiring us to rewrite parts of the MEDLINE section.

Cumulative Index to Nursing and Allied Health Literature

CINAHL is an important electronic database: It covers references to virtually all English-language nursing and allied health journals, as well as to books, dissertations, and selected conference proceedings in nursing and allied health fields. There are several versions of the CINAHL database (e.g., CINAHL, CINAHL Plus), each with somewhat different features relating to full text availability and journal coverage. All are offered through EBSCOhost.

The basic CINAHL database indexes material from nearly 3,000 journals dating from 1981, and contains more than 1 million records. In addition to providing information for locating references (i.e., author, title, journal, year of publication, volume, and page numbers), CINAHL provides abstracts of most citations. Supplementary information, such as names of data collection instruments, is available for many records. CINAHL can be accessed through CINAHL (*www.ebscohost.com/cinahl/*) or through institutional libraries. We illustrate features of CINAHL, but note that some may be labeled differently at your institution.

At the outset, you might begin with a "basic search" by simply entering keywords or phrases relevant to your primary question. In the basic search screen, you could limit your search in a number of ways, for example, by limiting the records retrieved to those with certain features (e.g., only ones with abstracts or only those in journals with peer review), to specific publication dates (e.g., only those from 2005 to the present), or to those coded as being in a particular subset (e.g., nursing). The basic search

screen also allows you to expand your search by clicking an option labeled "Apply related words."

As an example, suppose we were interested in recent research on nurses' pain management for children. If we searched for *pain*, we would get nearly 20,000 records. Searching for *pain AND child* AND nurs** would bring the number down to about 2,000. (In CINAHL, an asterisk is the truncation symbol and a question mark is the wildcard). We could pare the number down to about 300 in a basic search by limiting the search to articles with abstracts published in nursing journals after 2004.

The advanced search mode in CINAHL permits even more fine-tuning. For example, we could stipulate that we wanted only research articles published in English. These restrictions, which take only seconds to execute, would get us down to a more manageable number of records (130) that could be searched more carefully for relevance. The advanced search mode offers many additional search options that should be more fully explored.

The full records for the 130 references would then be displayed on the monitor in a Results List. The Results List has sidebar options that allow you to narrow your search even farther, if desired. From the Results List, we could place promising references into a folder for later scrutiny, or we could immediately retrieve and print full bibliographic information for records of interest. An example of an abridged CINAHL record entry for a study identified through the search on children's pain is presented in Figure 5.2. The record begins with the article title, the authors' names and affiliation, and source. The source indicates the following:

- Name of the journal (*Pediatric Nursing*)
- Year and month of publication (2008 Jul–Aug)
- Volume (34)
- Issue (4)
- Page numbers (297–397)
- Number of cited references (40)

The record also shows the major and minor CINAHL subject headings that were coded for this study. Any of these headings could have been used to retrieve this reference. Note that the subject headings include substantive codes such as *Pain – Nursing*, and also methodologic codes (e.g., *Questionnaires*) and sample characteristic codes (e.g., *Child*). Next, the abstract for the study is shown. Based on the abstract, we would decide whether this reference was pertinent. Additional information on the record includes the journal subset, special interest category, instrumentation, and (if relevant) funding for the study. Each entry shows an accession number that is the unique identifier for each record in the database, as well as other identifying numbers.

An important feature of CINAHL and other databases is that it allows you to easily find other relevant references once a good one has been found. For example, in Figure 5.2 you can see that the record offers many embedded links on which you can click. For example, you could click on any of the authors' names to see if they have published other related articles. You could also click on any of the subject headings to track down other leads. There is also a link in each record called *Cited References*. By clicking this link, the entire reference list for the record (i.e., all the references cited in the article) would be retrieved, and you could then examine any of the citations. Finally, there is a sidebar link in each record called "*Find similar results*," which would retrieve additional records for articles with a similar focus.

In CINAHL, you can also explore the structure of the database's thesaurus to get additional leads for searching. The tool bar at the top of the screen has a tab called *CINAHL Headings*. When you click on this tab and enter a term in the "Browse" field, you can enter a term of interest and select one of three options: Term Begins With, Term Contains, or Relevance Ranked (which is the default). For example, if we entered *pain* and then clicked on Browse, we would be shown the 52 relevant subject headings relating to pain. We could then search the database for any of the listed subject headings. Also, many terms have an "Explode" option, which allows you to create a search query in which headings are exploded to retrieve all references indexed to that term.

Title:	Nurse characteristics and inferences about children's pain
Authors:	Griffin RA; Polit DF; Byrne MW
Affiliation:	Boston College, School of Nursing, Chestnut Hill, MA
Source:	Pediatric Nursing (PEDIATR ***NURS***), 2008 Jul–Aug; 34(4): 297–307 (40 ref)
Publication Type:	journal article – CEU, exam questions, ***research***, tables/charts
Language:	English
Major Subjects:	Child, Hospitalized Nurse Attitudes – Evaluation Pain – Nursing Pain – Therapy – In Infancy and Childhood Pediatric Nursing
Minor Subjects:	Analysis of Variance; Child; Cross Sectional Studies; Demography; Descriptive Statistics; Female; Mail; Male; Multiple Regression; Post Hoc Analysis; Questionnaires; Random Sample; Scales; Survey Research; T-Tests; United States; Vignettes; Visual Analog Scaling
Abstract:	The purpose of this study was to describe pediatric ***nurses***' projected responses to ***children's pain*** as described in vignettes of hospitalized ***children*** and to explore ***nurse*** characteristics that might influence those responses. A survey was mailed to a national random sample of 700 RNs, and 334 ***nurses*** responded. The survey included case reports of three hospitalized school-aged ***children*** experiencing ***pain***. ***Nurses*** were asked to rate their perceptions of the ***children's pain*** levels and to indicate how much analgesia they would recommend. Contrary to earlier studies, in response to the scenarios, ***nurses*** in this sample perceived high levels of ***pain***, said they would administer doses of analgesia close to the maximum prescribed by physicians, and recommended an array of non-pharmacologic methods to treat ***pain***. Variation in ***pain*** perceptions and decisions was not related to key personal and professional characteristics of the ***nurses***, including their education level, race/ethnicity, age, years of clinical experience, and receipt of continuing education about ***pain***. Findings from this large national study suggest that most ***nurses*** would make appropriate decisions relating to the treatment of ***children's pain***, perhaps reflecting changes in the emphasis on ***pain*** management.
Journal Subset:	Core nursing; Nursing; Peer reviewed; USA
Special Interest:	Pain and Pain Management; Pediatric Care
Instrumentation:	FACES pain scale (FPS)
Accession No.	2010006653

FIGURE 5.2 Example of a record from a CINAHL search.

CINAHL can also be used to pursue descendancy searches. In the Results List, there is a notation for each record entry for the number of times the article was cited in the CINAHL database. Clicking on the link would show the full list of articles that had cited this study.

TIP: The Institute for Scientific Information (ISI) maintains a multidisciplinary resource called the Web of Knowledge, which offers searching opportunities in several bibliographic databases. The Web of Knowledge is widely used for its citation feature, which can be helpful in applying a descendancy strategy, using a link labeled "Cited Reference."

The MEDLINE Database

The MEDLINE database was developed by the U.S. National Library of Medicine (NLM), and is widely recognized as the premier source for bibliographic coverage of the biomedical literature. MEDLINE covers about 5,000 medical, nursing, and health journals published in about 70 countries and contains more than 15 million records dating back to the mid 1960s. In 1999, abstracts of reviews from the Cochrane Collaboration became available through MEDLINE.

The MEDLINE database can be accessed online through a commercial vendor such as Ovid, but this

database can be accessed for free through the PubMed website (*http://www.ncbi.nlm.nih.gov/PubMed*). This means that anyone, anywhere in the world, with Internet access can search for journal articles, and thus, PubMed is a lifelong resource regardless of your institutional affiliation. PubMed has an excellent tutorial.

On the Home page of PubMed, you can launch a basic search that looks for your keyword in text fields of the record. As you begin to enter your keyword (or a key phrase) in the search box, automatic suggestions will display, and you can click on the one that is the best match.

MEDLINE uses a controlled vocabulary called MeSH (Medical Subject Headings) to index articles. MeSH provides a consistent way to retrieve information that may use different terms for the same concepts. You can learn about relevant MeSH terms by clicking on the "MeSH database" link on the home page (under "More Resources"). If, for example, we searched the MeSH database for "pain," we would find that Pain is a MeSH subject heading (a definition is provided) and there are 39 additional related categories—for example, "pain measurement" and "somatoform disorders." MeSH subject headings may overlap with, but are not identical to, subject headings used in CINAHL.

If you begin using your own keyword in a basic search, you can see how your term mapped onto MeSH terms by scrolling down and looking in the right-hand panel for a section labeled "Search Details." For example, if we entered the keyword "children" in the search field of the initial screen, Search Details would show us that PubMed searched for all references that have "child" or "children" in text fields of the database record, *and* it also searched for all references that had been coded "child" as a subject heading, because "child" is a MeSH subject heading. When you initiate a search, PubMed offers an "Also Try" feature (also in the right panel) that suggests other terms to enter in the search field (e.g., pain children).

If we did a PubMed search of MEDLINE similar to the one we described earlier for CINAHL, we would find that a simple search for *pain* would yield about 420,000 records, and *pain AND child* AND nurs** would yield nearly 2,500. We can place restrictions on the search by clicking the blue "Limits" link right above the search box. Limits include date (e.g., published in the last 2 years), language (e.g., English), journal subset (e.g., Nursing journals), and text options (e.g., only those with abstracts). If we limited our search to entries with abstracts, written in English, published within the past 5 years, and coded in the Nursing subset, the search would yield about 300 citations. This PubMed search yielded more references than the CINAHL search, but we were not able to limit the search to research reports: PubMed does not have a generic category that distinguishes all research articles from nonresearch articles. Further options for building the search are available by clicking the "Advanced Search" link, which is directly to the right of the "Limits" link.

Figure 5.3 shows the full citation for the same study we located earlier in CINAHL (Figure 5.2). Beneath the abstract, when you click on "MeSH Terms" the display presents all of the MeSH terms that were used for this particular study, and also any "Substances." As you can see, the MeSH terms are quite different from the subject headings for the same reference in CINAHL. As with CINAHL, you can click on highlighted record entries (author names and MeSH terms) for possible leads. You can also click on a link labeled "LinkOut," which provides more resources for the article. In this example, the link tells us that there are three full text sources for this study: EBSCO, Ovid, and ProQuest (not shown in Figure 5.3).

In the right panel of the screen for PubMed records there is a list of "Related Articles," which is a useful feature once you have found a study that is a good exemplar of the evidence for which you are looking. Further down in the right panel, PubMed provides a list of any articles in the MEDLINE database that had cited this study, which is useful for a descendancy search.

TIP: Searching for qualitative studies can pose special challenges. Walters and colleagues (2006) described how they developed optimal search strategies for qualitative studies in the EMBASE database, and Wilczynski and colleagues (2007) offered advice for searching in CINAHL. Flemming and Briggs (2006) compared three alternative strategies for finding qualitative research.

Pediatr Nurs. 2008 Jul–Aug;34(4):297–305.

Nurse characteristics and inferences about children's pain.

Griffin RA, **Polit DF**, **Byrne MW**.

Boston College, School of Nursing, Chestnut Hill, MA, USA.

The purpose of this study was to describe pediatric nurses' projected responses to children's pain as described in vignettes of hospitalized children and to explore nurse characteristics that might influence those responses. A survey was mailed to a national random sample of 700 RNs, and 334 nurses responded. The survey included case reports of three hospitalized school-aged children experiencing pain. Nurses were asked to rate their perceptions of the children's pain levels and to indicate how much analgesia they would recommend. Contrary to earlier studies, in response to the scenarios, nurses in this sample perceived high levels of pain, said they would administer doses of analgesia close to the maximum prescribed by physicians, and recommended an array of nonpharmacologic methods to treat pain. Variation in pain perceptions and decisions was not related to key personal and professional characteristics of the nurses, including their education level, race/ethnicity, age, years of clinical experience, and receipt of continuing education about pain. Findings from this large national study suggest that most nurses would make appropriate decisions relating to the treatment of children's pain, perhaps reflecting changes in the emphasis on pain management.
PMID: 18814563 [PubMed - indexed for MEDLINE]

MeSH Terms:
Analgesics, Opioid/administration & dosage
Child
Cross-Sectional Studies
Female
Health Care Surveys
Health Knowledge, Attitudes, Practice*
Humans
Male
Middle Aged
Pain/drug therapy
Pain/nursing*
Pain Measurement*
United States

Substances: Analgesics, Opioid

FIGURE 5.3 Example of a record from a PubMed search.

Screening and Gathering References

References that have been identified through a literature search need to be screened. One screen is a practical one: Is the reference accessible? For example, some references may be written in a language you do not read, or published in a journal that you cannot retrieve. A second screen is relevance, which you can usually infer by reading the abstract. If an abstract is unavailable, you will need to guess about relevance based on the title. When screening an article, keep in mind that some of the articles judged to be not relevant for your primary question may be appropriate for a secondary question. A third screening criterion may be the study's methodologic quality—i.e., the quality of evidence the study yields, a topic discussed in a later section.

We strongly urge you to obtain full copies of relevant studies rather than taking notes. It is often necessary to reread an article or to get further details about a study, which can easily be done if you have a copy. Online retrieval of full text articles has increasingly become possible. An article that is not directly available online through your institution can be retrieved through a commercial vendor, by photocopying it from a hardcopy journal, or by requesting a copy from the lead author via e-mail communication.

Each obtained article should be filed in a manner that permits easy access. Some authors (Garrard, 2006) advocate a chronological filing method (e.g., by date of publication), but we think that alphabetical filing (using last name of the first author) is easier.

Documentation in Literature Retrieval

If your goal is to "own" the literature, you will be using a variety of databases, keywords, subject headings, and strategies in your effort to pursue all possible leads. As you meander through the complex world of research information, you will likely lose track of your efforts if you do not document your actions from the outset.

It is highly advisable to maintain a notebook (or computer database program) to record your search strategies and search results. You should make note of information such as databases searched; limits put on your search; specific keywords, subject headings, or authors used to direct the search; combining strategies adopted; studies used to inaugurate a "Related Articles" or "descendancy" search; websites visited; links pursued; authors contacted to request further information or copies of articles not readily available; and any other information that would help you keep track of what you have done. Part of your strategy usually can be documented by printing your search history from electronic databases.

By documenting your actions, you will be able to conduct a more efficient search—that is, you will not inadvertently duplicate a strategy you have already pursued. Documentation will also help you to assess what else needs to be tried—where to go next in your search. Finally, documenting your efforts is a step in ensuring that your literature review is reproducible.

TIP: The Toolkit section of the accompanying *Resource Manual* offers a template for documenting certain types of information during a literature search. The template, as a Word document, can easily be augmented and adapted.

ABSTRACTING AND RECORDING INFORMATION

Tracking down relevant research on a topic is only the beginning of doing a literature review. Once you have a stack of useful articles, you need to develop a strategy for making sense of the information in them. If a literature review is fairly simple, it may be sufficient to jot down notes about key features of the studies under review and to use these notes as the basis for your analysis. However, literature reviews are often complex—for example, there may be dozens of studies, or study findings may vary. In such situations, it is useful to adopt a formal system of recording key information about each study. We describe two mechanisms for doing this, formal protocols and matrices. First, though, we discuss the advantages of developing a coding scheme.

Coding the Studies

Reviewers who undertake systematic reviews often develop extensive coding systems to support statistical analyses. Coding may not be necessary in less formal reviews, but we do think that coding can be useful, so we offer some simple suggestions and an example.

To develop a coding scheme, you will need to read at least a subset of studies and look for opportunities to categorize information. One approach is to code for key variables or themes. Let us take the example we have used in this chapter, the relationship between nurses' characteristics (the independent variable) on the one hand and nurses' responses to children's pain (the dependent variable) on the other. By perusing the articles we retrieved, we find that several nurse characteristics have been studied—for example, their age, gender, clinical experience, and so on. We can assign codes to each characteristic. Now let us consider the dependent variable, nurses' responses to children's pain. We find that some studies have focused on nurses' perceptions of children's pain, others have examined nurses' use of analgesia, and so on. These different outcomes can also be coded. An

BOX 5.1 Codes for Results Matrix/Coding in Margins

CODES FOR NURSE CHARACTERISTICS (INDEPENDENT VARIABLES)

1. Age
2. Gender
3. Education
4. Years of clinical experience
5. Race/ethnicity
6. Personal experience with pain
7. Nurse practitioner status

CODES FOR RESPONSES TO CHILDREN'S PAIN (DEPENDENT VARIABLES)

a. Perceptions of children's pain
b. Pain treatment (use of analgesia)
c. Pain treatment (use of nonpharmacologic methods)
d. Other (e.g., perceived barriers to optimal pain management)

example of a simple coding scheme is presented in Box 5.1.

The codes can then be applied to the studies. You can record these codes in a protocol or matrices (which we discuss next), but you should also note the codes in the margins of the articles themselves, so you can easily find the information. Figure 5.4, which presents an excerpt from the results of a study by Vincent and Denyes (2004), shows marginal coding of key variables.

Coding can be a useful organizational tool even when a review is focused. For example, if our research question was about nurses' use of nonpharmacologic methods of pain treatment (i.e., not about use of analgesics or about pain perceptions), the outcome categories could be specific nonpharmacologic approaches, such as distraction, guided imagery, massage, and so on. The point is to organize information in a way that facilitates retrieval and analysis.

Literature Review Protocols

One method of organizing information from research articles is to use a formal protocol. Protocols are a means of recording various aspects of a study

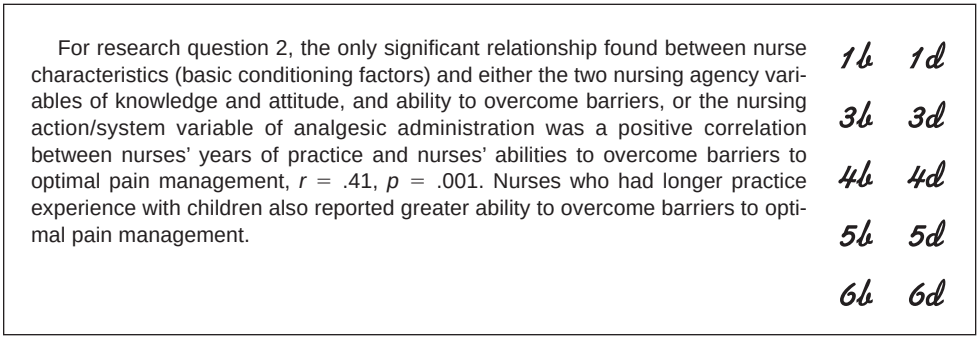


FIGURE 5.4 Coded excerpt from Results section. From Vincent, C. V., & Denyes, M. J. [2004]. Relieving children's pain: Nurses' abilities and analgesic administration practices. *Journal of Pediatric Nursing, 19*[1], 40–50.

systematically, including the full citation, theoretical foundations, methodologic features, findings, and conclusions. Evaluative information (e.g., your assessment of the study's strengths and weaknesses) can also be noted.

There is no fixed format for such a protocol—you must decide what elements are important to record *consistently* across studies to help you organize and analyze information. The example in Figure 5.5 can be adapted to fit your needs. (Although many

Citation: Authors: _____
Title: _____
Journal: _____
Year: _____ Volume: _____ Issue: _____ Pages: _____

Type of Study: ☐ Quantitative ☐ Qualitative ☐ Mixed Method

Location/Setting: _____

Key concepts/Variables: Concepts: _____
Intervention/Independent Variable: _____
Dependent Variable: _____
Controlled Variable: _____

Framework/Theory: _____

Design Type: ☐ Experimental ☐ Quasi-experimental ☐ Nonexperimental
Specific Design: _____
Blinding? ☐ None ☐ Single:_____ ☐ Double _____
Descrip. of Intervention: _____
_____
_____
Comparison group(s): _____
☐ Cross-sectional ☐ Longitudinal/Prospective No. of data collection points: ___

Qual. Tradition: ☐ Grounded theory ☐ Phenomenology ☐ Ethnography ☐ Other: _____

Sample: Size:_____ Sampling method: _____
Sample characteristics: _____
_____

Data Sources: Type: ☐ Self-report ☐ Observational ☐ Biophysiologic ☐ Other _____
Description of measures: _____
_____
_____
Data Quality: _____
_____

Statistical Tests: Bivariate: ☐ *t*-test ☐ ANOVA ☐ Chi-square ☐ Pearson's *r* ☐ Other: _____
Multivar: ☐ Multiple Regression ☐ MANOVA ☐ Logistic Regression ☐ Other: ____

Findings/Effect Sizes/Themes _____
_____
_____
_____
_____
_____

Recommendations: _____
_____

Strengths: _____
_____

Weaknesses: _____

FIGURE 5.5 Example of a literature review protocol.

terms on this protocol may not be familiar to you yet, you will learn their meaning in later chapters.) If you developed a coding scheme, you can use the codes to record information about study variables rather than writing out their names. Once you have developed a draft protocol, you should pilot test it with several studies to make sure it is sufficiently comprehensive.

Literature Review Matrices

For traditional narrative reviews of the literature, we prefer using two-dimensional matrices to organize information, because matrices directly support a thematic analysis. The content of the matrices, and number of matrices, can vary. A matrix can be constructed in hand-written form, in a word processing table, or in a spreadsheet. One advantage of computer files is that the information in the matrices can then be manipulated and sorted (e.g., the matrix entries can be sorted chronologically, or by authors' last name). We present some basic ideas, but there is room for creativity in designing matrices to organize information.

We think three types of matrix are useful:

- A Methodologic Matrix, which organizes information to answer: How have researchers studied this research question?
- Results Matrices, which address: What have researchers *found*?
- An Evaluation Matrix, to answer: How much confidence do we have in the evidence?

A Methodologic Matrix is used to record key features of study methods. Each row is for a study, and columns are for the kinds of methodologic information you want to capture across studies. An abbreviated example of such a matrix for the question about nurses' characteristics in relation to response to children's pain is presented in Figure 5.6 (available in the Toolkit). This matrix only has six entries (other relevant studies were omitted to save space), yet it is clear that information arrayed in this fashion allows us to see patterns that might otherwise have gone unnoticed. For example, by looking down the columns, we can readily discern that the broad research question has attracted international interest, samples of convenience have predominated, and self-report methods of data collection are most often used. When such a matrix is completed for all studies, it is easy to draw conclusions about how research questions have been addressed.

To discern themes in the pattern of *results*, we recommend developing multiple Results Matrices. It is useful to have as many Results Matrices as there are codes for either the independent or dependent variables, whichever is greater. In our coding scheme in Box 5.1, there are 7 independent variables and 4 dependent variables, so we would have 7 Results Matrices, one for each independent variable. The matrix in Figure 5.7, for example, is for recording information for studies that examined *nurses' education* in relation to responses to children's pain. Other matrices would focus on nurses' age, years of experience, and so on. In each matrix, columns are used for dependent variables, and rows represent separate studies. Findings about the relationship between a particular independent variable and a particular dependent variable are noted in the cells. The cell entries can indicate more precisely how dependent variables were operationalized, the direction of any relationships, level of significance, or other types of statistical information. Although there are only four studies in this Results Matrix, we can detect some patterns: the evidence, although not consistent, mostly suggests that nurses' level of education is unrelated to their responses to children's pain. Older studies were more likely than recent ones to find that more education was associated with better pain management.

Care should be taken in abstracting results information. Researchers sometimes point out only the findings that are statistically significant. Take, for example, the coded paragraph in Figure 5.4. The researchers (Vincent & Denyes, 2004) only elaborated results about the relationship between the nurses' years of experience and their ability to overcome barriers to optimal pain management. However, as indicated in the entry in the Methodologic Matrix (see Figure 5.6), this study gathered and analyzed data about 5 nurse characteristics in relation to 2 pain management outcomes, and so

Authors	Pub Yr	Country	Dependent Variables	Independent Variables	Study Design	Sample Size	Sampling Method	Data Collection	Age of Children
Griffin et al.	2008	U.S.A.	Perception of child's pain, Use of analgesics, Use of nonpharmacologic methods	Nurses' age, clinical experience, education, nurse practitioner status	Cross-sectional, correlational	332 nurses, national sample	Random	Self-report questionnaire	8–10
Twycross	2007	U.K.	Pain management practices	Knowledge of pain management	Cross-sectional, correlational	13 nurses, 1 surgical ward	Convenience	Observation, self-report	0–16
Vincent & Denyes	2004	U.S.A.	Use of analgesics, Perceived barriers to optimal pain management	Nurses' age, race, clinical experience, education, pain experience	Cross-sectional, correlational	67 nurses from 7 hospital units	Convenience	Observation, self-report questionnaire	3–17
Polkki et al.	2001	Finland	Nurses' use of nonpharmacologic methods	Nurses' age, education, clinical experience, # own kids	Cross-sectional, correlational	162 nurses from 5 hospitals	Convenience	Self-report questionnaire	8–12
Hamers et al.	1997	Netherlands	Assessments of child's pain, Confidence in assessment, Use of analgesics	Level of experience in pediatric nursing	Cross-sectional, correlational	695 nurses	Convenience	Video, vignette, self-reports	5–10
Margolius et al.	1995	U.S.A.	Perceptions of child's pain, Perceived adequacy of pain management	Nurses' education, age, years of nursing experience	Cross-sectional, correlational	228 nurses, 1 pediatric setting	Convenience	Self-report questionnaire	NA

FIGURE 5.6 Example of a methodologic matrix for recording key methodologic features of studies for a literature review: nurse characteristics and management of children's pain.

Independent Variable: *Nurses' Education* (Code 3)

Authors	Pub Year	DV.a Pain Perceptions	DV.b Use of Analgesics	DV.c Use of Nonpharmacologics	DV.d Other
Griffin et al.	2008	Rating of child's pain: no significant relationship	Amount used within PRN: no significant relationship	Number of nonpharmacologic strategies used: no significant relationship	—
Vincent & Denyes	2004	—	Percent of prescribed medications administered: no significant relation	—	Perceived barriers to optimal pain management: no significant relationship
Polkki et al.	2001	—	—	Use of nonpharmacologic methods higher in those with more education ($p < .05$)	—
Margolius et al.	1995	Perception of children's pain: higher with more education ($p < .05$)	—	—	—

FIGURE 5.7 Example of a results matrix for recording key findings for a literature review: nurses' education and management of children's pain.

there are 10 codes in the margin of Figure 5.4. Thus, although nothing in the paragraph mentions nurses' education, we have entered "no significant relationship" in two cells of the Results Matrix in Figure 5.7 because the paragraph implies that all relationships, except one, were nonsignificant.

TIP: Results matrices can also be used for qualitative studies. Instead of columns for independent or dependent variables, columns can be used to record themes, concepts, or categories.

CRITIQUING STUDIES AND EVALUATING THE EVIDENCE

In drawing conclusions about a body of research, reviewers must record not only factual information about studies—methodologic features and findings—but must also make judgments about the worth of the evidence. This section discusses issues relating to research critiques.

Research Critiques of Individual Studies

A research **critique** is a careful appraisal of the strengths and weaknesses of a study. A good critique objectively identifies areas of adequacy and inadequacy. Although our emphasis in this chapter is on the evaluation of a body of research evidence for a literature review, we pause to offer advice about other types of critiques.

Many critiques focus on a single study rather than on aggregated evidence. For example, most journals that publish research articles have a policy of soliciting critiques by two or more peer reviewers who prepare written critiques and make a recommendation about whether or not to publish the report. Peer reviewers' critiques typically are brief and focus on key substantive and methodologic issues.

Students taking a research course may be asked to critique a study, to document their mastery of methodologic concepts. Such critiques usually are expected to be comprehensive, encompassing various dimensions of a report. This might include substantive and theoretical aspects, ethical issues, methodologic decisions, interpretation, and the report's organization and presentation. The purpose of such thorough critique is to cultivate critical thinking, to induce students to use and document newly acquired research skills, and to prepare students for a professional nursing career in which evaluating research will almost surely play a role. Writing research critiques is an important first step on the path to developing an evidence-based practice.

TIP: When doing a research critique, you should read the article you are critiquing at least twice because the first step in preparing a critique is to understand what the report is saying. We encourage you to write in the margins of the article and to circle keywords.

We provide support for such comprehensive critiques of individual studies in several ways. First, detailed critiquing suggestions corresponding to chapter content are included in most chapters. Second, we offer an abbreviated set of key critiquing guidelines for quantitative and qualitative reports here in this chapter, in Boxes 5.2 and 5.3, respectively. Finally, it is always illuminating to have a good model, and so Appendices H and I of the accompanying *Resource Manual* include completed comprehensive research critiques of a quantitative and qualitative study (the studies themselves are printed in their entirety as well).

TIP: The guidelines in Boxes 5.2 and 5.3, which are available in the Toolkit of the accompanying *Resource Manual*, can be used to critique the quantitative and qualitative components of mixed methods studies that combine the two approaches (see Chapter 25). In addition, the questions in Box 25.1 should be addressed for a comprehensive critique of mixed methods studies.

The guidelines in Boxes 5.2 and 5.3 are organized according to the structure of most research articles—Abstract, Introduction, Method, Results, and Discussion. The second column lists key critiquing questions that have broad applicability to

BOX 5.2 Guide to an Overall Critique of a Quantitative Research Report

Aspect of the Report	Critiquing Questions	Detailed Critiquing Guidelines
Title	• Is the title a good one, succinctly suggesting key variables and the study population?	
Abstract	• Does the abstract clearly and concisely summarize the main features of the report (problem, methods, results, conclusions)?	
Introduction		
Statement of the problem	• Is the problem stated unambiguously, and is it easy to identify? • Does the problem statement build a cogent, persuasive argument for the new study? • Does the problem have significance for nursing? • Is there a good match between the research problem and the paradigm and methods used? Is a quantitative approach appropriate?	Box 4.3, page 90
Hypotheses or research questions	• Are research questions and/or hypotheses explicitly stated? If not, is their absence justified? • Are questions and hypotheses appropriately worded, with clear specification of key variables and the study population? • Are the questions/hypotheses consistent with the literature review and the conceptual framework?	Box 4.3, page 90
Literature review	• Is the literature review up to date and based mainly on primary sources? • Does the review provide a state-of-the-art synthesis of evidence on the problem? • Does the literature review provide a sound basis for the new study?	Box 5.4, page 122
Conceptual/ theoretical framework	• Are key concepts adequately defined conceptually? • Is there a conceptual/theoretical framework, rationale, and/or map, and (if so) is it appropriate? If not, is the absence of one justified?	Box 6.3, page 145
Method		
Protection of human rights	• Were appropriate procedures used to safeguard the rights of study participants? Was the study externally reviewed by an IRB/ethics review board? • Was the study designed to minimize risks and maximize benefits to participants?	Box 7.3, page 170

BOX 5.2 Guide to an Overall Critique of a Quantitative Research Report (continued)

Aspect of the Report	Critiquing Questions	Detailed Critiquing Guidelines
Research design	• Was the most rigorous possible design used, given the study purpose? • Were appropriate comparisons made to enhance interpretability of the findings? • Was the number of data collection points appropriate? • Did the design minimize biases and threats to the internal, construct, and external validity of the study (e.g., was blinding used, was attrition minimized)?	Box 9.1, page 230; Box 10.1, page 254
Population and sample	• Is the population described? Is the sample described in sufficient detail? • Was the best possible sampling design used to enhance the sample's representativeness? Were sampling biases minimized? • Was the sample size adequate? Was a power analysis used to estimate sample size needs?	Box 12.1, page 289
Data collection and measurement	• Are the operational and conceptual definitions congruent? • Were key variables operationalized using the best possible method (e.g., interviews, observations, and so on) and with adequate justification? • Are specific instruments adequately described and were they good choices, given the study purpose, variables being studied, and the study population? • Does the report provide evidence that the data collection methods yielded data that were reliable and valid?	Box 13.1, page 309; Box 14.1, page 347
Procedures	• If there was an intervention, is it adequately described, and was it rigorously developed and implemented? Did most participants allocated to the intervention group actually receive it? Is there evidence of intervention fidelity? • Were data collected in a manner that minimized bias? Were the staff who collected data appropriately trained?	Box 9.1, page 230; Box 10.1, page 254
Results		
Data analysis	• Were analyses undertaken to address each research question or test each hypothesis? • Were appropriate statistical methods used, given the level of measurement of the variables, number of groups being compared, and assumptions of the tests? • Was the most powerful analytic method used (e.g., did the analysis help to control for confounding variables)? • Were Type I and Type II errors avoided or minimized? • In intervention studies, was an intention-to-treat analysis performed? • Were problems of missing values evaluated and adequately addressed?	Box 16.1, page 400; Box 17.1, page 429

(box continues on page 114)

BOX 5.2 Guide to an Overall Critique of a Quantitative Research Report (continued)

Aspect of the Report	Critiquing Questions	Detailed Critiquing Guidelines
Findings	• Is information about statistical significance presented? Is information about effect size and precision of estimates (confidence intervals) presented? • Are the findings adequately summarized, with good use of tables and figures? • Are findings reported in a manner that facilitates a meta-analysis, and with sufficient information needed for EBP?	Box 17.1, page 429; Box 28.1, page 687
Discussion		
Interpretation of the findings	• Are all major findings interpreted and discussed within the context of prior research and/or the study's conceptual framework? • Are causal inferences, if any, justified? • Are interpretations well-founded and consistent with the study's limitations? • Does the report address the issue of the generalizability of the findings?	Box 19.1, page 482
Implications/ recommendations	• Do the researchers discuss the implications of the study for clinical practice or further research—and are those implications reasonable and complete?	Box 19.1, page 482
Global Issues		
Presentation	• Is the report well-written, organized, and sufficiently detailed for critical analysis? • In intervention studies, is a CONSORT flow chart provided to show the flow of participants in the study? • Is the report written in a manner that makes the findings accessible to practicing nurses?	Box 28.2, page 698
Researcher credibility	• Do the researchers' clinical, substantive, or methodologic qualifications and experience enhance confidence in the findings and their interpretation?	
Summary assessment	• Despite any limitations, do the study findings appear to be valid—do you have confidence in the *truth* value of the results? • Does the study contribute any meaningful evidence that can be used in nursing practice or that is useful to the nursing discipline?	

quantitative and qualitative studies, and the third column has cross-references to the detailed guidelines in the various chapters of the book. Many critiquing questions are likely too difficult for you to answer at this point, but your methodologic and critiquing skills will develop as you progress through this book. We developed these guidelines based on our years of experience as researchers and research methodologists, but they do not represent a formal, rigorously developed set of questions that

BOX 5.3 Guide to an Overall Critique of a Qualitative Research Report

Aspect of the Report	Critiquing Questions	Detailed Critiquing Guidelines
Title	• Is the title a good one, suggesting the key phenomenon and the group or community under study?	
Abstract	• Does the abstract clearly and concisely summarize the main features of the report?	
Introduction		
Statement of the problem	• Is the problem stated unambiguously and is it easy to identify? • Does the problem statement build a cogent and persuasive argument for the new study? • Does the problem have significance for nursing? • Is there a good match between the research problem on the one hand and the paradigm, tradition, and methods on the other?	Box 4.3, page 90
Research questions	• Are research questions explicitly stated? If not, is their absence justified? • Are the questions consistent with the study's philosophical basis, underlying tradition, or ideological orientation?	Box 4.3, page 90
Literature review	• Does the report adequately summarize the existing body of knowledge related to the problem or phenomenon of interest? • Does the literature review provide a sound basis for the new study?	Box 5.4, page 122
Conceptual underpinnings	• Are key concepts adequately defined conceptually? • Is the philosophical basis, underlying tradition, conceptual framework, or ideological orientation made explicit and is it appropriate for the problem?	Box 6.3, page 145
Method		
Protection of participants' rights	• Were appropriate procedures used to safeguard the rights of study participants? Was the study subject to external review by an IRB/ethics review board? • Was the study designed to minimize risks and maximize benefits to participants?	Box 7.3, page 170
Research design and research tradition	• Is the identified research tradition (if any) congruent with the methods used to collect and analyze data? • Was an adequate amount of time spent in the field or with study participants? • Did the design unfold in the field, giving researchers opportunities to capitalize on early understandings? • Was there an adequate number of contacts with study participants?	Box 20.1, page 510

(box continues on page 116)

BOX 5.3 Guide to an Overall Critique of a Qualitative Research Report (continued)

Aspect of the Report	Critiquing Questions	Detailed Critiquing Guidelines
Sample and setting	• Was the group or population of interest adequately described? Were the setting and sample described in sufficient detail? • Was the approach used to recruit participants or gain access to the site productive and appropriate? • Was the best possible method of sampling used to enhance information richness and address the needs of the study? • Was the sample size adequate? Was saturation achieved?	Box 21.1, page 528
Data collection	• Were the methods of gathering data appropriate? Were data gathered through two or more methods to achieve triangulation? • Did the researcher ask the right questions or make the right observations, and were they recorded in an appropriate fashion? • Was a sufficient amount of data gathered? Were the data of sufficient depth and richness?	Box 22.1, page 548
Procedures	• Are data collection and recording procedures adequately described and do they appear appropriate? • Were data collected in a manner that minimized bias? Were the staff who collected data appropriately trained?	Box 22.1, page 548
Enhancement of trustworthiness	• Did the researchers use effective strategies to enhance the trustworthiness/integrity of the study, and was the description of those strategies adequate? • Were the methods used to enhance trustworthiness appropriate and sufficient? • Did the researcher document research procedures and decision processes sufficiently that findings are auditable and confirmable? • Is there evidence of researcher reflexivity? • Is there "thick description" of the context, participants, and findings, and was it at a sufficient level to support transferability?	Box 24.1, page 598; Table 24.1, page 587
Results		
Data analysis	• Are the data management and data analysis methods sufficiently described? • Was the data analysis strategy compatible with the research tradition and with the nature and type of data gathered? • Did the analysis yield an appropriate "product" (e.g., a theory, taxonomy, thematic pattern)? • Do the analytic procedures suggest the possibility of biases?	Box 23.1, page 559

BOX 5.3 Guide to an Overall Critique of a Qualitative Research Report (continued)

Aspect of the Report	Critiquing Questions	Detailed Critiquing Guidelines
Findings	• Are the findings effectively summarized, with good use of excerpts and supporting arguments? • Do the themes adequately capture the meaning of the data? Does it appear that the researcher satisfactorily conceptualized the themes or patterns in the data? • Does the analysis yield an insightful, provocative, authentic, and meaningful picture of the phenomenon under investigation?	Box 23.1, page 559
Theoretical integration	• Are the themes or patterns logically connected to each other to form a convincing and integrated whole? • Are figures, maps, or models used effectively to summarize conceptualizations? • If a conceptual framework or ideological orientation guided the study, are the themes or patterns linked to it in a cogent manner?	Box 23.1, page 559; Box 6.3, page 145
Discussion		
Interpretation of the findings	• Are the findings interpreted within an appropriate social or cultural context? • Are major findings interpreted and discussed within the context of prior studies? • Are the interpretations consistent with the study's limitations?	Box 23.1, page 559
Implications/ recommendations	• Do the researchers discuss the implications of the study for clinical practice or further inquiry—and are those implications reasonable and complete?	
Global Issues		
Presentation	• Is the report well written, organized, and sufficiently detailed for critical analysis? • Is the description of the methods, findings, and interpretations sufficiently rich and vivid?	Box 28.2, page 698
Researcher credibility	• Do the researchers' clinical, substantive, or methodologic qualifications and experience enhance confidence in the findings and their interpretation?	
Summary assessment	• Do the study findings appear to be trustworthy—do you have confidence in the *truth* value of the results? • Does the study contribute any meaningful evidence that can be used in nursing practice or that is useful to the nursing discipline?	

are appropriate for a formal systematic review. They should, however, facilitate beginning efforts to critically appraise nursing studies. (Some formal guidelines are referenced in Chapter 27).

A few comments about these guidelines are in order. First, the questions call for a yes or no answer (although for some, the answer may be "Yes, *but* . . ."). In all cases, the desirable answer is "yes." That is, a "no" suggests a possible limitation and a "yes" suggests a strength. Therefore, the more "yeses" a study gets, the stronger it is likely to be. These guidelines can thus cumulatively suggest a global assessment: a report with 25 "yeses" is likely to be superior to one with only 10. Not all "yeses" are equal, however. Some elements are more important in drawing conclusions about study rigor than others. For example, the inadequacy of the article's literature review is less damaging to the worth of the study's *evidence* than the use of a faulty design. In general, questions about methodologic decisions (i.e., the questions under "Method") and about the analysis are especially important in evaluating the study's evidence.

Although the questions in these boxes elicit yes or no responses, a comprehensive critique would need to do more than point out what the researchers did and did not do. Each relevant issue would need to be discussed and your criticism justified. For example, if you answered "no" to the question about whether the problem was easy to identify, you would need to describe your concerns and perhaps offer suggestions for improvement.

Our simplified critiquing guidelines have a number of shortcomings. In particular, they are generic despite the fact that critiquing cannot use a one-size-fits-all list of questions. Some critiquing questions that are relevant to, say, clinical trials do not fit into a set of general questions for all quantitative studies. Thus, you would need to use some judgment about whether the guidelines are sufficiently comprehensive for the type of study you are critiquing, and perhaps supplement them with the more detailed critiquing questions in each chapter of this book.

Finally, there are questions in these guidelines for which there are no objective answers. Even experts sometimes disagree about what are the best methodologic strategies for a study. Thus, you should not be afraid to express an evaluative opinion—but be sure that your comments have some basis in methodologic principles discussed in this book.

TIP: It is appropriate to assume the posture of a skeptic when you are critiquing a research article. Just as a careful clinician seeks evidence from research findings that certain practices are or are not effective, you as a reviewer should demand evidence from the article that the researchers' decisions and their conclusions were sound.

Evaluating a Body of Research

In reviewing the literature, you typically would not undertake a *comprehensive* critique of each study—but you would need to assess the quality of evidence in each study so that you could draw conclusions about the overall body of evidence. Critiques for a literature review tend to focus on methodologic aspects.

In systematic reviews, methodologic quality often plays a role in selecting studies because investigations judged to be of low quality are sometimes screened out from further consideration. Using methodologic quality as a screening criterion is controversial, however. Systematic reviews sometimes involve the use of a formal evaluation instrument that gives quantitative ratings to aspects of the study, so that appraisals across studies ("scores") can be compared. Methodologic screening and formal scoring instruments are described in Chapter 27.

In literature reviews for a new primary study, methodologic features of studies under review need to be assessed with an eye to answering a broad question: To what extent do the findings reflect the *truth* or, conversely, to what extent do biases undermine the believability of the findings? The "truth" is most likely to be revealed when researchers use powerful designs, good sampling plans, strong data collection instruments and procedures, and appropriate analyses.

Judgments about the rigor of studies under review can be entered in an Evaluation Matrix.

Authors	Year of Publication	Major Strengths	Major Weaknesses	Quality Score*
Vincent & Denyes	2004	• Measured actual use of analgesics, not self-report • Linkage to Orem's theory • Good descriptive info on knowledge, attitudes, and use of analgesics	• Small and unrepresentative sample ($N = 67$), strong likelihood of Type II error (questionable power analysis) • Weak design for studying Q1 (effect of knowledge on analgesic use, effect of analgesic use on actual pain); several internal validity threats • Possibility that nurses' behavior in administering analgesics was affected by knowing they were in a study	12
Study 2				
Study 3				

*The quality score is fictitious and is shown here to indicate that information of this type could be recorded in the evaluation matrix.

FIGURE 5.8 Example of an evaluation matrix for recording strengths and weaknesses of studies for a literature review: nurse characteristics and management of children's pain.

Alternatively, additional columns for evaluative information can be added to the Methodologic Matrix. The advantage of combining information in one matrix is that methodologic features and assessments about those features are in a single table. The disadvantage is that the matrix would have so many columns that it might be cumbersome. A simple Evaluation Matrix is presented in Figure 5.8, which provides space in the columns for noting major strengths and weaknesses for each study (the rows). If a "score" for overall quality is derived from a formal scoring instrument (e.g., by counting all the "yeses" from Boxes 5.2 or 5.3), this information can be added to the Evaluation Matrix.

ANALYZING AND SYNTHESIZING INFORMATION

Once all the relevant studies have been retrieved, read, abstracted, and critiqued, the information has to be analyzed and synthesized. As previously noted, doing a literature review is similar to doing a qualitative study, particularly with respect to the analysis of the "data" (i.e., information from the retrieved studies). In both, the focus is on identifying important *themes*.

A thematic analysis essentially involves detecting patterns and regularities, as well as inconsistencies. Several different types of themes can be identified, as described in Table 5.1. The reason we have recommended using various matrices should be clear from reading this list of possible themes: it is easier to discern patterns by reading down the columns of the matrices than by flipping through a stack of review protocols.

Clearly, it is not possible—even in lengthy free-standing reviews—to analyze all the themes we have identified. Reviewers have to make decisions about which patterns to pursue. In preparing a review as part of a new study, you would need to determine which pattern is of greatest relevance for developing an argument and providing a context for the new research.

TABLE 5.1 Thematic Possibilities for a Literature Review

TYPE OF THEME	QUESTIONS FOR THEMATIC ANALYSIS
Substantive	What is the pattern of evidence? How much evidence is there? How consistent is the body of evidence? How powerful are the observed effects? How persuasive is the evidence? What gaps are there in the body of evidence?
Theoretical	What theoretical or conceptual frameworks have been used to address the primary question—or has most research been atheoretical? How congruent are the theoretical frameworks? Do findings vary in relation to differences in frameworks?
Generalizability/ Transferability	To what types of people or settings do the findings apply? Do the findings vary for different types of people (e.g., men versus women) or setting (e.g., urban versus rural)?
Historical	Have there been substantive, theoretical, or methodologic trends over time? Is the evidence getting better? When was most of the research conducted?
Researcher	Who has been doing the research, in terms of discipline, specialty area, nationality, prominence, and so on? Has the research been developed within a systematic program of research?

PREPARING A WRITTEN LITERATURE REVIEW

Writing literature reviews can be challenging, especially when voluminous information must be condensed into a small number of pages, as is typical for a journal article or proposal. We offer a few suggestions, but acknowledge that skills in writing literature reviews develop over time.

Organizing the Review

Organization is crucial in a written review. Having an outline helps to structure the flow of presentation. If the review is complex, a written outline is recommended; a mental outline may suffice for simpler reviews. The outline should list the main topics or themes to be discussed, and indicate the order of presentation. The important point is to have a plan before starting to write so that the review has a coherent flow. The goal is to structure the review in such a way that the presentation is logical, demonstrates meaningful thematic integration, and leads to a conclusion about the state of evidence on the topic.

Writing a Literature Review

Although it is beyond the scope of this textbook to offer detailed guidance on writing research reviews, we offer a few comments on their content and style. Additional assistance is provided in books such as those by Fink (2009) and Galvan (2009).

Content of the Written Literature Review

A written research review should provide readers with an objective, organized synthesis of evidence

on a topic. A review should be neither a series of quotes nor a series of abstracts. The central tasks are to summarize and critically evaluate the overall evidence so as to reveal the current state of knowledge—not simply to describe what researchers have done.

Although key studies may be described in some detail, it is not necessary to provide particulars for every reference, especially when there are page constraints. Studies with comparable findings often can be summarized together.

Example of grouped studies: Considine and McGillivray (2010) summarized several studies as follows in their introduction to a study of emergency nursing care for acute stroke: "Although the use of thrombolysis as a treatment option for acute stroke is discussed in most stroke guidelines…, most current evidence does not support the use of thrombolysis in acute ischaemic stroke beyond three hours (Hacke et al., 1995; Clarke et al., 1999, 2000; Kothari et al., 2001; National Stroke Foundation, 2003) to 4–5 hours after symptom onset (Haack et al., 2008, Wahlgren et al., 2008)."

The literature should be summarized in your own words. The review should demonstrate that you have considered the cumulative worth of the body of research. Stringing together quotes from various documents fails to show that previous research has been assimilated and understood.

The review should be objective, to the extent possible. Studies that are at odds with your hypotheses should not be omitted, and the review should not ignore a study because its findings contradict other studies. Inconsistent results should be analyzed and the supporting evidence evaluated objectively.

A literature review typically concludes with a concise summary of current evidence on the topic and gaps in the evidence. If the review is conducted for a new study, this critical summary should demonstrate the need for the research and should clarify the basis for any hypotheses.

TIP: As you progress through this book, you will acquire proficiency in critically evaluating studies. We hope you will understand the *mechanics* of doing a review after reading this chapter, but we do not expect you to be able to write a state-of-the-art review until you have gained more skills in research methods.

Style of a Research Review

Students preparing their first written research review often face stylistic challenges. In particular, students sometimes accept research findings uncritically, perhaps reflecting a common misunderstanding about the conclusiveness of research. You should keep in mind that hypotheses cannot be proved or disproved by empirical testing, and no research question can be definitely answered in a single study. This does not mean that research evidence should be ignored. The problem is partly semantic: hypotheses are not proved, they are *supported* by research findings. Research reviews should be written in a style that suggests tentativeness.

TIP: When describing study findings, you can use phrases indicating tentativeness of the results, such as the following:

- Several studies have *found* . . .
- Findings thus far *suggest* . . .
- Results from a landmark study *indicated* . . .
- The data *supported* the hypothesis . . .
- There *appears* to be strong evidence that . . .

A related stylistic problem is the interjection of opinions into the review. The review should include opinions sparingly, if at all, and should be explicit about their source. Reviewers' own opinions do not belong in a review, except for assessments of study quality.

The left-hand column of Table 5.2 presents several examples of stylistic flaws for a review. The right-hand column offers suggestions for rewordings that are more acceptable for a research literature review. Many alternative wordings are possible.

TABLE 5.2 Examples of Stylistic Difficulties for Research Literature Reviews

PROBLEMATIC STYLE OR WORDING	IMPROVED STYLE OR WORDING
Women who do not participate in childbirth preparation classes manifest a high degree of anxiety during labor.	*Studies have found that* women who participate in childbirth preparation classes *tend to* manifest less anxiety than those who do not (Franck, 2011; Kim, 2010; Yepsen, 2011).
Studies have proved that doctors and nurses do not fully understand the psychobiologic dynamics of recovery from a myocardial infarction.	Studies by Fortune (2010) and Crampton (2011) *suggest that many* doctors and nurses do not fully understand the psychobiologic dynamics of recovery from a myocardial infarction.
Attitudes cannot be changed quickly.	Attitudes *have been found to be* relatively stable, enduring attributes that do not change quickly (Nicolet, 2010; Brusser & Lace, 2011)
It is known that uncertainty engenders stress.	*According to* Dr. A. Cassard (2011), an expert on stress and anxiety, uncertainty is a stressor.

Note: Italicized words in the improved version indicate key alternations.

CRITIQUING RESEARCH LITERATURE REVIEWS

It is often difficult to critique a research review because the author is almost invariably more knowledgeable about the topic than the readers. It is thus not usually possible to judge whether the author has included all relevant literature and has adequately summarized evidence on that topic. Many aspects of a review, however, are amenable to evaluation by readers who are not experts on the topic. Some suggestions for critiquing written research reviews are presented in Box 5.4. When a review is published as

BOX 5.4 Guidelines for Critiquing Literature Reviews

1. Is the review thorough—does it include all of the major studies on the topic? Does it include recent research? Are studies from other related disciplines included, if appropriate?
2. Does the review rely on appropriate materials (e.g., mainly on primary source research articles)?
3. Is the review merely a summary of existing work, or does it critically appraise and compare key studies? Does the review identify important gaps in the literature?
4. Is the review well organized? Is the development of ideas clear?
5. Does the review use appropriate language, suggesting the tentativeness of prior findings? Is the review objective? Does the author paraphrase, or is there an over reliance on quotes from original sources?
6. If the review is part of a research report for a new study, does the review support the need for the study?
7. If it is a review designed to summarize evidence for clinical practice, does the review draw reasonable conclusions about practice implications?

a stand-alone article, it should include information to help readers evaluate the reviewer's search strategies, as discussed in Chapter 27.

In assessing a literature review, the key question is whether it summarizes the current state of research evidence adequately. If the review is written as part of an original research report, an equally important question is whether the review lays a solid foundation for the new study.

RESEARCH EXAMPLES OF LITERATURE REVIEWS

The best way to learn about the style, content, and organization of a research literature review is to read reviews in nursing journals. We present excerpts from two reviews here and urge you to read others on a topic of interest to you.*

Literature Review from a Quantitative Research Report

Study: Accuracy of vaginal symptom self-diagnosis algorithms for deployed military women (Ryan-Wenger et al, 2010)

Statement of Purpose: The major purpose of this study was to evaluate the accuracy of a prototype of the Women in the Military Self-Diagnosis (WMSD) kit for the diagnosis of vaginal symptoms. Another aim was to predict potential self-medication omission and commission error rates.

Literature Review (Excerpt): "Deployment settings are typically austere, characterized by extreme temperatures, primitive sanitary conditions, and limited hygiene and laundry facilities. These factors increase military women's risk for vaginitis. . . . Ryan-Wenger and Lowe (2000) surveyed 1,537 military women about their symptoms of genitourinary infections and healthcare experiences in their home duty stations and during deployment. Of the 841 women who had been deployed, 87% ($n = 732$) reported that they experienced vaginal symptoms such as itching, discharge, or foul odor at some time during deployment. Because of these symptoms, nearly half the women (48%) noted a decrease in the quality of their work performance and 24% lost from a few hours to more than a day of work time. . . . In focus groups conducted by DACOWITS [Defense Department Advisory Committee on Women in the Services], in our survey, and in a phenomenological study of soldier care, women evaluated deployment healthcare services for women as inadequate, citing lack of confidence in the knowledge and skills of the provider, lack of privacy, and lack of confidentiality (DACOWITS, 2007; Jennings, 2005; Ryan-Wenger & Lowe, 2000). . . . We proposed that a viable solution to the problem is a field-expedient kit for self-diagnosis and self-treatment of common genitourinary symptoms. . . .

Despite . . . diagnostic standards, studies show that clinicians often misdiagnose vaginal infections. For example, in one study, 197 vaginal samples were analyzed by culture, Gram stain, microscopy, and DNA hybridization with Affirm TM VPIII to derive a diagnosis of BV [bacterial vaginosis], TV [trichomonas vaginitis], and/or CV [candida vaginitis] (Schweiertz et al., 2006). Compared with laboratory diagnoses, physicians misdiagnosed CV in 77.1% of 109 cases, BV in 61.3% of 80 cases, and 87.5% of eight mixed infections. One reason for such high levels of inaccuracy is that many providers do not use the common office-based tests that are recommended to achieve a diagnosis. This point is illustrated by a study of diagnostic procedures used by physicians with 52 women who made 150 visits to a vaginitis clinic (Wiesenfeld & Macio, 1999). Microscopic assessment was done in 63% of the visits, and whiff and pH tests were conducted in only 3% of visits. In another study, 556 nurse practitioners and 608 physicians reported their diagnostic practices on a Web-based survey (Anderson & Karasz, 2005). An average of 79% of these providers indicated that they 'often or always' examined women with vaginal symptoms, 47% conducted whiff tests, and only 33.5% conducted pH tests on vaginal fluid" (pp. 2–4).

Literature Review from a Qualitative Research Report

Study: Young people's experience of living with ulcerative colitis and an ostomy (Savard & Woodgate, 2009)

Statement of Purpose: The purpose of this study was to understand the lived experiences of young adults with inflammatory bowel disease and an ostomy.

Literature Review (Excerpt): "Ulcerative colitis (UC) and Crohn's disease are collectively referred to as inflammatory bowel disease (IBD). . . . Approximately 25% of all new Crohn's disease cases and between 15% and 40% of all new UC cases are diagnosed in individuals younger

*Consult the full research reports for references cited in these excerpted literature reviews.

than 20 years of age (Kim & Ferry, 2004; Rayhorn, 2001). Individuals with IBD experience a range of symptoms including abdominal pain, cramping, and loose stools (Listrom & Holt, 2004; Pearson, 2004; Rayhorn, 2001; Veronesi, 2003). Some individuals may at some point during their illness require surgery, resulting in an ostomy (Reynaud & Meeker, 2002).

Although there has been discussion in the literature about what it is like to have IBD with or without an ostomy, young people (i.e., adolescents and young adults) have rarely been asked about their experiences (Daniel, 2001; Decker, 2000). Of the research done on young people, a lack of consensus remains as to how IBD affects this population socially and psychologically. Some studies reveal that IBD has negative psychological effects such as alienation, reduced living space, feelings of helplessness, self-blame, depression, and anxiety (Brydolf & Segesten, 1996; Daniel, 2001; Dudley-Brown, 1996; Mackner & Crandall, 2006; Wood et al., 1987), whereas others reveal that people with IBD cope well and are psychologically healthy (Joachim & Milne, 1987; Mackner & Crandall, 2005).

Studies carried out on individuals living with ostomies reveal that they face many lifestyle challenges that include physical and psychological adjustments (Manderson, 2005; Reynaud & Meeker, 2002; Rheaume & Gooding, 1991; Slater, 1992). Others have found that individuals with a temporary or permanent stoma perceive negative body image feelings and express difficulties in coming to terms with having the stoma (Black, 2004; Casati et al., 2000; Junkin & Beitz, 2005; . . .), especially the young population (O'Brien, 1999; Willis, 1998). . . .

A limitation of the work to date is that it has mainly been approached from a quantitative paradigm, and hence is not focused on capturing the meanings that young people ascribe to their experience. The literature review revealed four qualitative studies, two Swedish and two Canadian, that focus on the lived experienced of young individuals with IBD (Brydoff & Segeston, 1996; Daniel, 2001; Nicholas et al., 2007; Reichenberg et al., 2007). Although involving young people from different countries, common findings included the young people experiencing a reduced living space because of their dependency on needing to be near a toilet, feelings of embarrassment, a loss of control, and alienation from oneself and from others. . . . In summary, there is a need for more qualitative research that is directed at gaining understanding about the lived experiences of young people living with IBD and an ostomy" (pp. 33–34).

SUMMARY POINTS

- A research **literature review** is a written summary of evidence on a research problem.
- The major steps in preparing a written research review include formulating a question, devising a search strategy, conducting a search, retrieving relevant sources, abstracting information, critiquing studies, analyzing aggregated information, and preparing a written synthesis.
- Study findings are the major focus of research reviews. Information in nonresearch references—for example, opinion articles, case reports—may broaden understanding of a research problem, but has limited utility in research reviews.
- A **primary source** is the original description of a study prepared by the researcher who conducted it; a **secondary source** is a description of the study by a person unconnected with it. Literature reviews should be based on primary source material.
- Strategies for finding studies on a topic include the use of bibliographic tools, but also include the **ancestry approach** (tracking down earlier studies cited in a reference list of a report) and the **descendancy approach** (using a pivotal study to search forward to subsequent studies that cited it.)
- An important method for locating references is an electronic search of bibliographic databases. For nurses, the CINAHL and MEDLINE databases are especially useful.
- In searching a database, users can perform a **keyword search** that looks for searcher-specified terms in text fields of a database record (or that maps keywords onto the database's subject codes) or can search according to **subject heading** codes themselves.
- References must be screened for relevance, and then pertinent information must be abstracted for analysis. Formal review protocols and matrices facilitate abstraction.
- Matrices (two-dimensional arrays) are a convenient means of abstracting and organizing information for a literature review. A reviewer

might use a Methodologic Matrix to record methodologic features of a set of studies, a set of Results Matrices to record research findings, and an Evaluation Matrix to record quality assessment information. The use of such matrices facilitates thematic analysis of the retrieved information.

- A research **critique** is a careful appraisal of a study's strengths and weaknesses. Critiques for a research review tend to focus on the methodologic aspects of a set of studies. Critiques of individual studies tend to be more comprehensive.
- The analysis of information from a literature search involves the identification of important themes—regularities (and inconsistencies) in the information. Themes can take many forms, including substantive, methodologic, and theoretical themes.
- In preparing a written review, it is important to organize materials logically, preferably using an outline. The written review should not be a succession of quotes or abstracts. The reviewers' role is to describe study findings, the dependability of the evidence, evidence gaps, and (in the context of a new study) contributions that the new study would make.

STUDY ACTIVITIES

Chapter 5 of the *Resource Manual for Nursing Research: Generating and Assessing Evidence for Nursing Practice, 9th ed.*, offers study suggestions for reinforcing concepts presented in this chapter. In addition, the following questions can be addressed in classroom or online discussions:

1. Suppose you were planning to study the relationship between chronic transfusion therapy and quality of life in adolescents with sickle cell disease. Identify 5 to 10 keywords that could be used to search for relevant studies, and compare them with those found by other students.
2. Suppose you were studying factors affecting the discharge of chronic psychiatric patients. Obtain references for 5 studies for this topic, and compare them with those of other students.
3. Carefully examine Figures 5.6 and 5.7 and see how many themes you can identify. Also, see how many incongruities there are among studies in the matrixes (i.e., the absence of consistent themes).

STUDIES CITED IN CHAPTER 5

Considine, J., & McGillivray, B. (2010). An evidence-based practice approach to improving nursing care of acute stroke in an Australian Emergency Department. *Journal of Clinical Nursing, 19*, 138–144.

Griffin, R. A., Polit, D. F., & Byrne, M. W. (2008). Nurse characteristics and inferences about children's pain. *Pediatric Nursing, 34*, 297–307.

Hamers, J. P., van den Hout, M. A., Halfens, R. J., Abu-Saad, H. H., & Heijlties, A. E. (1997). Differences in pain assessment and decisions regarding the administration of analgesics between novices, intermediates and experts in pediatric nursing. *International Journal of Nursing Studies, 34*, 325–334.

Margolius, F. R., Hudson, K. A., & Miche, Y. (1995). Beliefs and perceptions about children in pain: A survey. *Pediatric Nursing, 21*, 111–115.

Murray, M., Fiset, V., Young, S., & Kryworuchko, J. (2009). Where the dying live: Review of determinants of place of end-of-life cancer care. *Oncology Nursing Forum, 36*, 69–77.

Polkki T., Vehvilainen-Julkunen K., & Pietila, A. M. (2001). Nonpharmacological methods in relieving children's postoperative pain: A survey on hospital nurses in Finland. *Journal of Advanced Nursing, 34*, 483–492.

Ryan-Wenger, N., Neal, J., Jones, A., & Lower, N. (2010). Accuracy of vaginal symptom self-diagnosis algorithms for deployed military women. *Nursing Research, 59*, 2–10.

Savard, J., & Woodgate, R. (2009). Young people's experience of living with ulcerative colitis and an ostomy. *Gastroenterology Nursing, 32*, 33–41.

Stajduhar, K., Allan, D., Cohen, S., & Heyland, D. (2008). Preferences for location of death of seriously ill hospitalized patients. *Palliative Medicine, 22*, 85–88.

Twycross, A. (2007). What is the impact of theoretical knowledge of children's nurses' post-operative pain management practices? *Nurse Education Today, 27*, 697–707.

Vincent, C. V., & Denyes, M. J. (2004). Relieving children's pain: Nurses' abilities and analgesic administration practices. *Journal of Pediatric Nursing, 19*, 40–50.

Methodologic and nonresearch references cited in this chapter can be found in a separate section at the end of the book.

9 Quantitative Research Design

GENERAL DESIGN ISSUES

Part 3 of this book (Chapters 9 through 19) focuses on methods of doing quantitative research.

This chapter describes options for designing quantitative studies. We begin by discussing several broad issues.

Causality

As noted in Chapter 2, several broad categories of research questions are relevant to evidence-based nursing practice—questions about interventions, diagnosis and assessment, prognosis, etiology and harm, and meaning or process (Table 2.1). Questions about meaning or process call for a qualitative approach, which we describe in Chapter 20. Questions about diagnosis or assessment, as well as questions about the status quo of health-related situations, are typically descriptive. Many research questions, however, are about *causes* and *effects:*

- Does a telephone therapy intervention for patients diagnosed with prostate cancer *cause* improvements in their decision-making skills? (intervention question)
- Do birthweights under 1,500 grams *cause* developmental delays in children? (prognosis question)
- Does cigarette smoking *cause* lung cancer? (etiology/harm question)

Although causality is a hotly debated philosophical issue, we all understand the general concept of a **cause**. For example, we understand that failure to sleep *causes* fatigue and that high-caloric intake *causes* weight gain.

Most phenomena have multiple causes. Weight gain, for example, can be the effect of high-caloric consumption, but other factors also cause weight gain. Causes of health-related phenomena usually are not *deterministic*, but rather *probabilistic*—that is, the causes increase the probability that an effect will occur. For example, there is ample evidence that smoking is a cause of lung cancer, but not everyone who smokes develops lung cancer, and not everyone with lung cancer was a smoker.

The Counterfactual Model

While it might be easy to grasp what researchers have in mind when they talk about a *cause*, what exactly is an **effect**? Shadish and colleagues (2002), who wrote a widely acclaimed book on research design and causal inference, explained that a good way to grasp the meaning of an effect is by

conceptualizing a counterfactual. In a research context, a **counterfactual** is what would have happened *to the same people* exposed to a causal factor if they *simultaneously* were *not* exposed to the causal factor. An effect represents the difference between what actually did happen with the exposure and what would have happened without it. This counterfactual model is an idealized conception that can never be realized, but it is a good model to keep in mind in designing a study to provide cause-and-effect evidence. As Shadish and colleagues (2002) noted, "A central task for all cause-probing research is to create reasonable approximations to this physically impossible counterfactual" (p. 5).

Criteria for Causality

Several writers have proposed criteria for establishing a cause-and-effect relationship. Lazarsfeld (1955), reflecting ideas of John Stuart Mill, identified three criteria for causality. The first is *temporal*: A cause must precede an effect in time. If we were testing the hypothesis that aspertame causes fetal abnormalities, it would be necessary to demonstrate that the abnormalities did not develop before the mothers' exposure to aspertame. The second requirement is that there be an *empirical relationship* between the presumed cause and the presumed effect. In the aspertame example, we would have to find an association between aspertame consumption and fetal abnormalities, that is, that a higher percentage of aspertame users than nonusers had infants with fetal abnormalities. The final criterion for inferring causality is that the relationship cannot be explained as being *caused by a third variable*. Suppose, for instance, that people who used aspertame tended also to drink more coffee than nonusers of aspertame. There would then be a possibility that any relationship between maternal aspertame use and fetal abnormalities reflects an underlying causal relationship between a substance in coffee and the abnormalities.

Additional criteria were proposed by Bradford-Hill (1971) as part of the discussion about the causal link between smoking and lung cancer. Two of Bradford-Hill's criteria foreshadow the importance of meta-analyses, techniques for which had not been fully developed when the criteria were proposed. The criterion of *coherence* involves having similar evidence from multiple sources, and the criterion of *consistency* involves having similar levels of statistical relationship in several studies. Another important criterion is *biologic plausibility*, that is, evidence from laboratory or basic physiologic studies that a causal pathway is credible.

Researchers investigating causal relationships must provide persuasive evidence about these criteria through their study design. Some designs are better at revealing cause-and-effect relationships than others, but not all research questions can be answered using the strongest designs because of ethical or practical constraints. Much of this chapter concerns designs for illuminating causal relationships.

Design Terminology

It is easy to get confused about terms used for research designs because there is inconsistency among writers. Moreover, design terms used by medical and epidemiologic researchers are usually different from those used by social scientists. Many early nurse researchers got their research training in social science fields such as psychology or sociology before doctoral-level training became available in schools of nursing, and so social scientific design terms have predominated in the nursing research literature.

Nurses interested in establishing an evidence-based practice must to be able to understand studies from many disciplines. We use both medical and social science terms in this book, although the latter predominate. Table 9.1 provides a list of several design terms used by social scientists and the corresponding terms used by medical researchers.

EXPERIMENTAL DESIGN

A basic distinction in quantitative research design is between experimental and nonexperimental research. In an **experiment** (or **randomized controlled trial, RCT**), researchers are active agents,

TABLE 9.1 Research Design Terminology in the Social Scientific and Medical Literature

SOCIAL SCIENTIFIC TERM	MEDICAL RESEARCH TERM
Experiment, true experiment, experimental study	Randomized controlled trial, randomized clinical trial, RCT
Quasi-experiment, quasi-experimental study	Controlled trial, controlled trial without randomization
Nonexperimental study, correlational study	Observational study
Retrospective study	Case-control study
Prospective nonexperimental study	Cohort study
Group or condition (e.g., experimental or control group/condition)	Group or arm (e.g., intervention or control arm)
Experimental group	Treatment or intervention group

not passive observers. Early physical scientists learned that although pure observation of phenomena is valuable, complexities occurring in nature often made it difficult to understand relationships. This problem was addressed by isolating phenomena in a laboratory and controlling the conditions under which they occurred. Procedures developed by physical scientists were profitably adopted by biologists during the 19th century, resulting in many achievements in physiology and medicine. The 20th century witnessed the increased use of experimental methods by researchers interested in human behavior.

The controlled experiment is considered to be the gold standard for yielding reliable evidence about causes and effects. Experimenters can be relatively confident in the genuineness of causal relationships because they are observed under controlled conditions and typically meet the criteria for establishing causality. As we pointed out in Chapter 4, hypotheses are never proved or disproved by scientific methods, but true experiments offer the most convincing evidence about the effect one variable has on another.

A true experimental or RCT design is characterized by the following properties:

- *Manipulation:* The researcher *does* something to at least some participants—that is, there is some type of intervention.
- *Control:* The researcher introduces controls over the experimental situation, including devising an approximation of a counterfactual—usually, a control group that does not receive the intervention.
- *Randomization*: The researcher assigns participants to a control or experimental condition on a random basis.

Design Features of True Experiments

Researchers have many options in designing an experiment. We begin by discussing several features of experimental designs.

Manipulation: The Experimental Intervention

Manipulation involves *doing* something to study participants. Experimenters manipulate the independent variable by administering a **treatment** (**intervention**) to some people and withholding it from others, or administering a different treatment. Experimenters deliberately *vary* the independent

variable (the presumed cause) and observe the effect on the outcome.

For example, suppose we hypothesized that gentle massage is an effective pain relief strategy for nursing home residents. The independent variable, receipt of gentle massage, can be manipulated by giving some patients the massage intervention and withholding it from others. We would then compare pain levels (the dependent variable) in the two groups to see if differences in receipt of the intervention resulted in differences in average pain levels.

In designing RCTs, researchers make many decisions about what the experimental condition entails, and these decisions can affect the conclusions. To get a fair test, the intervention should be appropriate to the problem, consistent with a theoretical rationale, and of sufficient intensity and duration that effects might reasonably be expected. The full nature of the intervention must be delineated in formal protocols that spell out exactly what the treatment is. Among the questions researchers need to address are the following:

- What *is* the intervention, and how does it differ from usual methods of care?
- What specific procedures are to be used with those receiving the intervention?
- What is the dosage or intensity of the intervention?
- Over how long a period will the intervention be administered, how frequently will it be administered, and when will the treatment begin (e.g., 2 hours after surgery)?
- Who will administer the intervention? What are their credentials, and what type of special training will they receive?
- Under what conditions will the intervention be withdrawn or altered?

The goal in most RCTs is to have an identical intervention for all people in the treatment group. For example, in most drug studies, those in the experimental group are given the exact same ingredient, in the same dose, administered in exactly the same manner—all according to well-articulated protocols. There is, however, growing interest in **patient-centered interventions** or **PCIs** (Lauver et al., 2002). The purpose of PCIs is to enhance treatment efficacy by taking people's characteristics or needs into account. In tailored interventions, each person receives an intervention customized to certain characteristics, such as demographic characteristics (e.g., gender), cognitive factors (e.g., reading level), or affective factors (e.g., motivation). Interventions based on the Transtheoretical (stages of change) Model (Chapter 6) usually are PCIs, because the intervention is tailored to fit people's readiness to change their behavior. There is some evidence that tailored interventions are more effective than standardized interventions (e.g., Lauver et al., 2003). More research in this area is needed, however, and such research is likely to play an important role in our current evidence-based practice environment in which there is a strong interest in understanding not only *what* works, but what works for *whom.*

TIP: Although PCIs are not universally standardized, they are typically administered according to well-defined procedures and guidelines, and the intervention agents are carefully trained in making decisions about who should get what type of treatment.

Manipulation: The Control Condition

Evidence about relationships requires making at least one comparison. If we were to supplement the diet of premature infants with a special nutrient for 2 weeks, their weight at the end of 2 weeks would tell us nothing about treatment effectiveness. At a bare minimum, we would need to compare posttreatment weight with pretreatment weight to determine if, at least, their weight had increased. But, let us assume that we find an average weight gain of 1 pound. Does this gain support the conclusion that the nutrition supplement (the independent variable) caused weight gain (the dependent variable)? No, it does not. Babies normally gain weight as they mature. Without a control group—a group that does *not* receive the supplement—it is impossible to separate the effects of maturation from those of the treatment.

The term **control group** refers to a group of participants whose performance on an outcome is used to evaluate that of the treatment group on the same outcome. As noted in Table 9.1, researchers with

training from a social science tradition use the term "group" or "condition" (e.g., the experimental group or the control condition), but medical researchers often use the term "arm," as in the intervention arm or the control arm of the study.

The control condition is a proxy for an ideal counterfactual. Researchers have choices about what to use as the counterfactual. Their decision is sometimes based on theoretical or substantive grounds, but may be driven by practical or ethical concerns. In some research, control group members receive no treatment at all—they are merely observed with respect to performance on the outcome. This type of control condition is not usually feasible in nursing research. For example, if we wanted to evaluate the effectiveness of a nursing intervention for hospital patients, we would not devise an RCT in which patients in the control group received no nursing care at all. Among the possibilities for the counterfactual are the following:

1. An alternative intervention; for example, participants could receive two different types of distraction as alternative therapies for pain.
2. A **placebo** or pseudointervention presumed to have no therapeutic value; for example, in studies of the effectiveness of drugs, some patients get the experimental drug and others get an innocuous substance. Placebos are used to control for the nonpharmaceutical effects of drugs, such as the attention being paid to participants. (There can, however, be **placebo effects**—changes in the dependent variable attributable to the placebo condition—because of participants' expectations of benefits or harms).

Example of a placebo control group: In a study of the effect of sucrose on infant pain responses during routine immunizations, Hatfield (2008) randomly assigned infants to groups administered either a sucrose solution or sterile water.

3. Standard methods of care—the usual procedures used to care for patients. This is the most typical control condition in nursing studies.
4. Different doses or intensities of treatment wherein all participants get some type of intervention, but the experimental group gets an intervention that is richer, more intense, or longer. This approach is attractive when there is a desire to analyze **dose-response effects**, that is, to test whether larger doses are associated with larger benefits, or whether a smaller (and perhaps less costly or burdensome) dose would suffice.

Example of different dose groups: Martinez and colleagues (2009) used an experimental design to test the relative effect of three "doses" of a walking intervention for patients with peripheral arterial disease. Participants were randomly assigned to a walking program lasting 2 to 9 weeks, 10 to 14 weeks, or 15 to 94 weeks.

5. **Wait-list control group**, with delayed treatment; the control group eventually receives the full experimental intervention, after all research outcomes are assessed.

Example of a wait-list control group: Heidrich and colleagues (2009) assessed the efficacy of an individualized intervention to improve symptom management in older breast cancer survivors. In one of their pilot studies, participants were assigned at random to the treatment condition or to a wait-list control group.

Methodologically, the best test is between two conditions that are as different as possible, as when the experimental group gets a strong treatment and the control group gets no treatment. Ethically, the most appealing counterfactual is probably the delay of treatment approach (number 5), which may be hard to do pragmatically. Testing two competing interventions (number 1) also has ethical appeal, but the risk is that the results will be inconclusive because it is difficult to detect differential effects if both interventions are at least moderately effective.

Some researchers combine two or more comparison strategies. For example, they might test two alternative treatments (option 1) against a placebo (option 3). Another option is to compare an intervention, a placebo, and no treatment. The use of multiple comparison groups is often attractive but, of course, adds to the cost and complexity of the study.

Example of a three-group design: Nikolajsen and colleagues (2009) randomly assigned patients undergoing placement of a femoral nerve block to one of three groups: two alternative intervention groups (audiovisual stimulation versus audio stimulation) or a "usual care" control group. Differences in pain were then assessed.

Sometimes researchers include an **attention control group** when they want to rule out the possibility that intervention effects are caused by the special attention given to those receiving the intervention, rather than by the actual treatment content. The idea is to separate the "active ingredients" of the treatment from the "inactive ingredients" of special attention.

Example of an attention control group: Seers and colleagues (2008) studied the effectiveness of relaxation for reducing postoperative pain and anxiety in orthopedic surgery patients. The design involved four groups—total body relaxation, jaw relaxation, attention control, and usual care control. Those in the attention control group received usual care, plus extra attention by being asked to describe what they do, feel, and think when they are in pain.

The control group decision should be based on an underlying conceptualization of how the intervention might "cause" the intended effect, and should also reflect consideration of what it is that needs to be controlled. For example, if attention control groups are being considered, there should be an underlying conceptualization of the construct of "attention" (Gross, 2005).

Whatever decision is made about a control group strategy, researchers need to be as careful in spelling out the counterfactual as in delineating the intervention. In research reports, researchers sometimes say that the control group got "usual methods of care" without explaining what that condition was and how different it was from the intervention being tested. In drawing on an evidence base for practice, nurses need to understand exactly what happened to study participants in different conditions. Barkauskas and colleagues (2005) and Shadish and colleagues (2002) offer useful advice about developing a control group strategy.

Randomization

Randomization (also called **random assignment** or **random allocation**) involves assigning participants to treatment conditions at random. *Random* means that everyone has an equal chance of being assigned to any group. If people are placed in groups randomly, there is no systematic bias in the groups with respect to preintervention attributes that could affect outcome variables.

Randomization Principles. The overall purpose of random assignment is to approximate the ideal—but impossible—counterfactual of having the same people in multiple treatment groups simultaneously. For example, suppose we wanted to study the effectiveness of a contraceptive counseling program for multiparous women who have just given birth. Two groups of women are included—one will be counseled and the other will not. Women in the sample are likely to differ from one another in many ways, such as age, marital status, financial situation, and the like. Any of these characteristics could affect a woman's diligence in practicing contraception, independent of whether she receives counseling. We need to have the "counsel" and "no counsel" groups equal with respect to these confounding characteristics to assess the impact of counseling on subsequent pregnancies. A counterfactual group needs to be equivalent, to the fullest extent possible, to the intervention group. Random assignment of people to one group or the other is designed to perform this equalization function. One method might be to flip a coin (more elaborate procedures are discussed later). If the coin comes up "heads," a participant would be assigned to one group; if it comes up "tails," she would be assigned to the other group.

Although randomization is the preferred method for equalizing groups, there is no *guarantee* that the groups will be equal. As an example, suppose the study sample involves 10 women who have given birth to 4 or more children. Five of the 10 women are aged 35 years or older, and the remaining 5 are younger than age 35. We would expect random assignment to result in two or three women from the two age ranges in each group. But suppose that, by chance, the older five women all ended up in the counseling group.

These women, who are nearing the end of childbearing years, have a lower likelihood of conceiving. Thus, follow-up of their subsequent childbearing might suggest that the counseling program was effective in reducing subsequent pregnancies; yet, a higher birth rate in the control group may reflect age and fecundity differences, not lack of exposure to counseling.

Despite this possibility, randomization is the most trustworthy method of equalizing groups. Unusual or deviant assignments such as this one are rare, and the likelihood of getting markedly unequal groups is reduced as the sample size increases.

You may wonder why we do not consciously control characteristics that are likely to affect the outcome through matching (Chapter 8). For example, if matching were used in the contraceptive counseling study, we could ensure that if there were a married, 38-year-old woman with six children in the experimental group, there would be a married, 38-year-old woman with six children in the control group. There are two problems with matching, however. First, to match effectively, we must know the characteristics that are likely to affect the outcome, but this knowledge is not always available. Second, even if we knew the relevant traits, the complications of matching on more than two or three characteristics simultaneously are prohibitive. With random assignment, *all* personal characteristics—age, income, intelligence, religiosity, and so on—are likely to be equally distributed in all groups. Over the long run, the groups tend to be counterbalanced with respect to an infinite number of biologic, psychological, economic, and social traits.

Basic Randomization. To demonstrate how random assignment is performed, we turn to another example. Suppose we were testing two alternative interventions to lower the anxiety of children who are about to undergo tonsillectomy. One intervention involves giving structured information about the surgical team's activities (procedural information); the other involves structured information about what the child will feel (sensation information). A third control group receives no special intervention. With a sample of 15 children, five will be randomly assigned to each group.

Researchers can use a **table of random numbers** to randomize. A small portion of such a table is shown in Table 9.2. In a table of random numbers, any digit from 0 to 9 is equally likely to follow any other digit. Going in any direction from any point in the table produces a random sequence.

In our example, we would number the 15 children from 1 to 15, as shown in column 2 of Table 9.3, and then draw numbers between 01 and 15 from the random number table. To find a random starting point, you can close your eyes and let your finger fall at some point on the table. For this example, assume that our starting point is at number 52, bolded in Table 9.2. We can move in any direction from that point, selecting numbers that fall between 01 and 15. Let us move to the right, looking at two-digit combinations. The number to the right of 52 is 06. The person whose number is 06, Nathan O., is assigned to group I. Moving along, the next number within our range is 11. (To find numbers in the desired range, we bypass numbers between 16 and 99.) Alaine J., whose number is 11, is also assigned to group I. The next three numbers are 01, 15, and 14. Thus, Kristina N., Chris L., and Paul M. are assigned to group I. The next five numbers between 01 and 15 in the table are used to assign five children to group II, and the remaining five are put into group III. Note that numbers that have already been used often reappear in the table before the task is completed. For example, the number 15 appeared four times during this randomization. This is normal because the numbers are random.

We can look at the three groups to see if they are equal for one readily discernible trait, gender. We started out with eight girls and seven boys. As Table 9.4 shows, randomization did a good job of allocating boys and girls about equally across the three groups. We must accept on faith the probability that other characteristics (e.g., race, age, initial anxiety) are also well distributed in the randomized groups. The larger the sample, the stronger the likelihood that the groups will be comparable across all factors that could affect the outcomes.

Researchers usually assign participants proportionately to groups being compared. For example, a sample of 300 participants in a 2-group design would generally be allocated 150 to the experimental

TABLE 9.2 Small Table of Random Digits

46	85	05	23	26	34	67	75	83	00	74	91	06	43	45
69	24	89	34	60	45	30	50	75	21	61	31	83	18	55
14	01	33	17	92	59	74	76	72	77	76	50	33	45	13
56	30	38	73	15	16	**52**	06	96	76	11	65	49	98	93
81	30	44	85	85	68	65	22	73	76	92	85	25	58	66
70	28	42	43	26	79	37	59	52	20	01	15	96	32	67
90	41	59	36	14	33	52	12	66	65	55	82	34	76	41
39	90	40	21	15	59	58	94	90	67	66	82	14	15	75
88	15	20	00	80	20	55	49	14	09	96	27	74	82	57
45	13	46	35	45	59	40	47	20	59	43	94	75	16	80
70	01	41	50	21	41	29	06	73	12	71	85	71	59	57
37	23	93	32	95	05	87	00	11	19	92	78	42	63	40
18	63	73	75	09	82	44	49	90	05	04	92	17	37	01
05	32	78	21	62	20	24	78	17	59	45	19	72	53	32
95	09	66	79	46	48	46	08	55	58	15	19	02	87	82
43	25	38	41	45	60	83	32	59	83	01	29	14	13	49
80	85	40	92	79	43	52	90	63	18	38	38	47	47	61
81	08	87	70	74	88	72	25	67	36	66	16	44	94	31
84	89	07	80	02	94	81	03	19	00	54	10	58	34	36

group and 150 to the control group. If there were 3 groups, there would be 100 per group. It is also possible (and sometimes desirable ethically) to have a different allocation. For example, if an especially promising treatment were developed, we could assign 200 to the treatment group and 100 to the control group. Such an allocation does, however, make it more difficult to detect treatment effects at statistically significant levels—or, to put it another way, the overall sample size must be larger to attain the same level of statistical reliability.

Computerized resources are available for free on the Internet to help with randomization. One such website is *www.randomizer.org*, which has a useful tutorial. Standard statistical software packages (e.g., SPSS or SAS) can also be used (see Shadish et al., 2002, p. 311). We also offer 2-digit and 3-digit random number tables in the Toolkit included with the accompanying *Resource Manual*.

TIP: There is considerable confusion—even in research methods textbooks—about random assignment versus random sampling. Randomization (random assignment) is a *signature* of an experimental design. If there is no random allocation of participants to conditions, then the design is not a true experiment. Random *sampling*, by contrast, is a method of selecting people for a study (see Chapter 12). Random sampling is *not* a signature of an experimental design. In fact, most RCTs do *not* involve random sampling.

Randomization Procedures. The success of randomization depends on two factors. First, the allocation process should be truly random. Second, there must be strict adherence to the randomization schedule. The latter can be achieved if the alloca-

TABLE 9.3 Example of Random Assignment Procedure

CHILD'S NAME	NUMBER	GROUP ASSIGNMENT
Kristina N.	01	I
Derek A.	02	III
Trinity A.	03	III
Lauren J.	04	II
Grace S.	05	II
Nathan O.	06	I
Norah J.	07	III
Thomas N.	08	III
Daniel B.	09	II
Rita T.	10	III
Alaine J.	11	I
Maren B.	12	II
Vadim B.	13	II
Paul M.	14	I
Chris L.	15	I

tion is unpredictable (for both participants and those enrolling them) and tamperproof. Random assignment should involve **allocation concealment** that prevents those who enroll participants from knowing upcoming assignments. Allocation concealment is intended to prevent biases that could stem from knowledge of allocations before assignments actually occur. To use an exaggerated example, if the person doing the enrollment knew that the next person enrolled would be assigned to a promising intervention, he or she might defer enrollment until a particularly needy patient came along. Allocation concealment can always be implemented, regardless of the intervention.

TABLE 9.4 Breakdown of the Gender Composition of the Three Groups

GENDER	GROUP I	GROUP II	GROUP III
Boys	3	2	2
Girls	2	3	3

Several methods have been devised to ensure allocation concealment, many of which involve developing a randomization schedule before the study begins. This is advantageous when people do not enter a study simultaneously, but rather on a *rolling enrollment* basis. In such situations, the sequence of allocation can be predetermined before enrollment. One widely used method is to have sequentially numbered, opaque sealed envelopes (SNOSE) containing assignment information. As each participant enters the study, he or she receives the next envelope in the sequence (for procedural suggestions, see Vickers, 2006, or Doig & Simpson, 2005). Envelope systems, however, can be subject to tampering (Vickers, 2006). A preferred method is to have treatment allocation information communicated to interventionists by a person unconnected with enrollment or treatment, by telephone or email. This person is trained to strictly follow the randomization schedule. In multisite trials, centralized randomization is strongly recommended.

TIP: Padhye and colleagues (2009) have described an easy-to-use spreadsheet method for randomization in small studies.

The timing of randomization is also important. Study eligibility—whether a person meets the criteria for inclusion—should be ascertained before randomization. If **baseline data** (preintervention data) are collected to measure key outcomes, this should occur before randomization to rule out any possibility that group assignment in itself might affect outcomes prior to treatment. Randomization should occur as closely as possible to the start of the intervention to maximize the likelihood that all randomized people will actually receive the condition to which they have been assigned. Figure 9.1 illustrates the sequence of steps that occurs in most RCTs, including the timing for obtaining informed consent.

Randomization Variants. In most cases, randomization involves the random assignment of individuals to different conditions. An alternative is **cluster randomization**, which involves randomly assigning *clusters* of people to different treatment groups (Christie et al.,

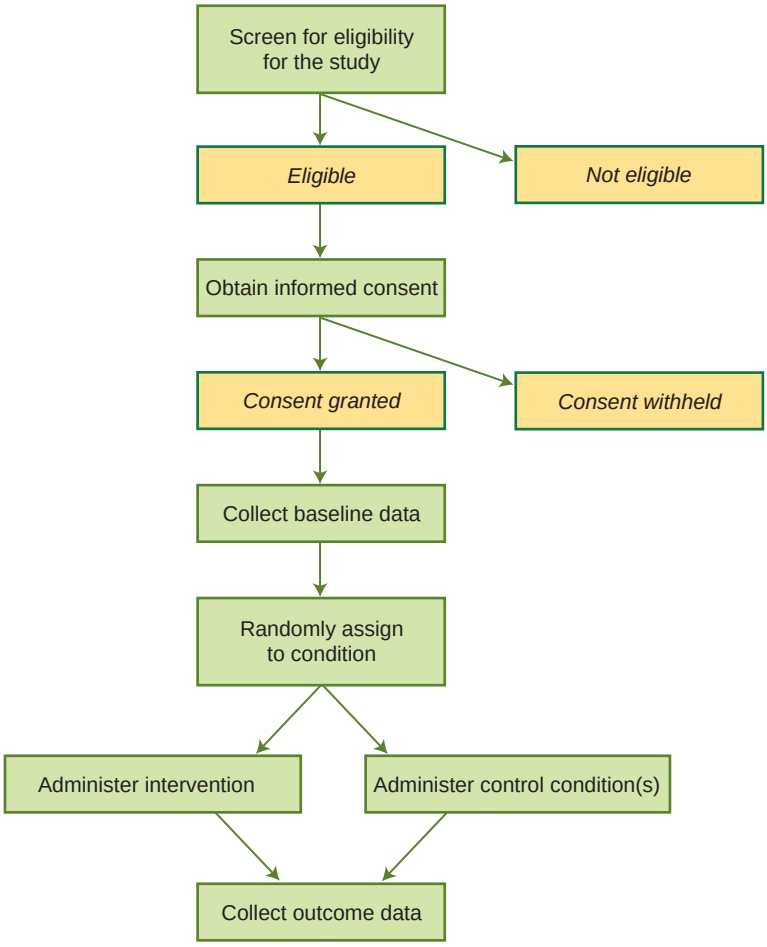


FIGURE 9.1 Sequence of steps in a conventional randomization design.

2009). Cluster randomization may enhance the feasibility of conducting an experiment. Groups of patients who enter a hospital unit at the same time, or patients at different sites, can be randomly assigned to a treatment condition as a unit—thus ruling out, in some situations, practical impediments to randomization. This approach also reduces the risk of **contamination of treatments**, that is, the co-mingling of people in the groups, which could cloud the results if they exchange information. The main disadvantages of cluster randomization are that the statistical analysis of data obtained through this approach is more complex, and sample size requirements are usually greater for a given level of accuracy. Moreover, the number of units being randomized must be fairly large for the randomization to be successful in equalizing across units. Cluster randomization can also complicate efforts at research synthesis using meta-analysis. Donner and Klar (2004) and Christie and colleagues (2009) offer useful discussions about planning a study with cluster randomization.

Example of cluster randomization: Huizing and colleagues (2009) tested an educational intervention to reduce the use of restraints in psychogeriatric nursing home wards. Fourteen wards were randomly assigned to receive the intervention or not. In all, 105 nursing home residents were included in the analyses.

Simple randomization is usually adequate for creating groups with comparable characteristics, but researchers sometimes take steps to ensure that subgroups of participants are allocated equally to conditions through **stratification**. For example, if a researcher stratified on the basis of gender, men and women would be randomly assigned to conditions separately, thus ensuring that both men and women received the intervention in the right proportions.

TIP: Sometimes stratification is called *blocking*, and the resulting design is called a **randomized block design**. This should not be confused with the design described next. When a cluster randomized design is used, it is almost always a good idea to first stratify units along a dimension of importance before randomizing.

Sometimes people are randomly assigned in blocks through **permuted block randomization**. Rather than having a randomization schedule for the entire sample, randomization occurs for blocks of participants—for example, 6 or 8 at a time. If the entire sample is randomly allocated to conditions, the first 5 or 6 people could be allocated to one or another condition, by chance alone. If allocation is done in randomly permuted blocks in randomly selected sizes, randomization within the small blocks would guarantee a balanced distribution across conditions while maintaing allocation concealment. Such a system is especially appropriate when enrollment occurs over a long period of time because the type of people enrolling might change—or the intervention itself might change due to improved proficiency in implementing it. The Toolkit in the *Resource Manual* offers guidance on block randomization.

Example of stratified, permuted block randomization: Lai and colleagues (2006) studied the effect of music during kangaroo care on maternal anxiety and infant response. Mother–infant dyads were randomly assigned to the treatment or control group using permuted block randomization, stratified on infant gender.

A controversial randomization variant is called **randomized consent** or a **Zelen design** after its originator (Zelen, 1979). Study participants sometimes have a preference about which condition they want. If randomization occurs *after* informed consent (as in Figure 9.1), people who are not assigned to their preferred condition may opt out of the study. Zelen proposed a simple solution: randomize first and *then* obtain consent, thus eliminating the possibility that the consent process will generate preferences. Those in the intervention group are then approached and offered the intervention, which they can accept or decline. If the control group condition is standard care, control group members may not even be asked for their consent, as they would not be getting anything different. The ethical controversies surrounding this form of randomization, as well as its merits and other limitations, have been described by Homer (2002).

Example of the Zelen design: Steiner and colleagues (2001) compared postacute intermediate care in a nurse-led unit versus conventional care on general medical wards in terms of such outcomes as patients' length of stay and mortality. The investigators, who used the Zelen design to randomize patients, argued that conventional randomization was distressful and confusing to many older patients.

Another method of addressing preferences is **partially randomized patient preference (PRPP)**, wherein all participants are asked preferences about treatment conditions. Only those without a strong preference are randomized, but all participants are followed up. Lambert and Wood (2000) outlined the benefits and problems of this approach.

Blinding or Masking

A rather charming (but problematic) quality of people is that they usually want things to turn out well. Researchers want their ideas to work, and they want their hypotheses supported. Participants often want to be helpful and also want to present themselves in a positive light. These tendencies can lead to biases because they can affect what participants do and say (and what researchers ask and perceive) in ways that distort the truth.

A procedure called **blinding** (or **masking**) is used in some RCTs to prevent biases stemming from *awareness*. Blinding involves concealing information from participants, data collectors, care

providers, intervention agents, or data analysts to enhance objectivity and minimize **expectation bias**. For example, if participants are not aware of whether they are getting an experimental drug or a placebo, then their outcomes cannot be influenced by their expectations of its efficacy. Blinding typically involves disguising or withholding information about participants' status in the study (e.g., whether they are in the experimental or control group), but can also involve withholding information about study hypotheses, baseline performance on outcomes, or preliminary study results.

The absence of blinding can result in different biases. **Performance bias** refers to systematic differences in the care provided to members of different groups of participants, apart from an intervention that is the focus of the inquiry. For example, participants in a "usual care" group may seek to obtain an innovative intervention elsewhere. Those delivering an intervention might treat participants in groups differently, apart from the intervention itself. Blinding of participants, and blinding agents delivering treatments, is used to avoid performance bias. **Detection** (or **ascertainment**) bias, which concerns systematic differences between groups in how outcome variables are measured, verified, or recorded, is addressed by blinding those who collect the outcome data or, in some cases, those who analyze them.

Unlike allocation concealment, blinding is not always possible. Drug studies often lend themselves to blinding, but many nursing interventions do not. For example, if the intervention were a smoking cessation program, participants would know that they were receiving the intervention, and the interventionist would be aware of who was in the program. However, it is usually possible, and desirable, to at least mask participants' treatment status from people collecting outcome data and from other clinicians providing normal care.

TIP: Although blinding is useful for minimizing bias, it may not be necessary if subjectivity and error risk are low. For example, participants' ratings of pain are subjective and susceptible to biases stemming from their own or data collectors' awareness of group status or study hypotheses. Hospital readmission and length of hospital stay, on the other hand, are variables less likely to be affected by people's awareness.

When blinding is not used, the study is an **open study**, in contrast to a **closed study** that results from masking. When blinding is used with only one group of people (e.g., study participants), it is sometimes described as a **single-blind study**. When it is possible to mask with two groups (e.g., those delivering an intervention and those receiving it), it is sometimes called **double-blind**, and when three groups are masked, it may be called **triple-blind**. However, recent guidelines have recommended that researchers not use these terms without explicitly stating which groups were blinded to avoid any ambiguity (Moher et al., 2010).

The term *blinding*, though widely used, has fallen into some disfavor because of possible pejorative connotations, and some organizations (e.g., the American Psychological Association) have recommended using masking instead. Medical researchers, however, appear to prefer *blinding* unless the people in the study have vision impairments (Schulz et al., 2002). Similarly, the vast majority of nurse researchers use the term *blinding* rather than *masking* (Polit et al., 2010).

Example of a single-blind experiment: Pölkki and colleagues (2008) tested an imagery-induced relaxation intervention to reduce postoperative pain in 8- to 12-year-old children. The nurse who collected the data did not know whether children were in the intervention group or the usual care control group.

Specific Experimental Designs

There are numerous experimental designs, including many that are not discussed in this book, such as *nested designs* and the *Solomon four-group design*. Some popular designs described in this section are summarized in Table 9.5. The second column (schematic diagram) depicts design notation from a classic monograph (Campbell & Stanley, 1963). In this notation, R means random assignment, O represents an observation (i.e., data collection on

TABLE 9.5 Design Alternatives: Selected Experimental (Randomized) Designs

TYPE OF DESIGN	SCHEMATIC DIAGRAM	SITUATIONS THAT ARE BEST SUITED TO THIS DESIGN	DRAWBACKS OF THIS DESIGN
1. Basic posttest-only design	R X O_1 R O_1 **or** R X_A O_1 R X_B O_1	When the outcome is not relevant until after the intervention is complete (e.g., length of stay in hospital)	Does not permit an evaluation of whether the two groups were comparable at the outset on the outcome of interest
2. Basic pretest–posttest design (with optional repeated follow-ups)	R O_1 X O_2 R O_1 O_2 R O_1 X O_2 O_3 O_4 R O_1 O_2 O_3 O_4	a. When the focus of the intervention is on change (e.g., behaviors, attitudes) b. When the researcher wants to assess both group differences (experimental comparison), and change within groups (quasi-experimental)	Sometimes the pretest itself can affect the outcomes of interest
3. Multiple intervention design	R O_1 X_A O_2 R O_1 X_B O_2 R O_1 O_2	Can be used to disentangle effects of different components of a complex intervention, or to test competing interventions	a. Requires larger sample than basic designs b. May be at risk to threats to statistical conclusion validity* if A and B are not very different (small effects)
4. Wait-list (delay of treatment) design	R O_1 X O_2 O_3 R O_1 O_2 X O_3	a. Attractive when there is patient preference for the innovative treatment b. Can strengthen inferences by virtue of replication aspect for the second group	a. Controls may drop out of study before they get deferred treatment b. Not suitable if key outcomes are measured long after treatment (e.g., mortality) or if there is an interest in assessing long-term effects (wait-list period is then too long)
5. Crossover design—participants serve as their own controls	R O_1 X_A O_2 X_B O_3 R O_1 X_B O_2 X_A O_3	a. Appropriate only if there is no expectation of carryover effects from one period to the next (effects should have rapid onset, short half-life) b. Useful when recruitment is difficult—smaller sample is needed; excellent for controlling confounding variables	a. Often cannot be assumed that there are no carryover effects b. If the first treatment received "fixes" a problem for participants, they may not remain in the study for the second one c. History threat* to validity a possibility
6. Factorial design	R O_1 X_{A1B1} O_2 R O_1 X_{A1B2} O_2 R O_1 X_{A2B1} O_2 R O_1 X_{A2B2} O_2	a. Efficient for testing two interventions simultaneously b. Can be useful in illuminating interaction effects, but most useful when strong synergistic/additive effects (or *no* interaction effects) are expected	Power needed to detect interactions could require larger sample size than when testing each intervention separately

KEY: R = Randomization
X = Intervention (X_A = one treatment, X_B = alternative treatment, dose, etc.)
O = Observation or measurement of the dependent variable/outcome

*Validity threats are discussed in Chapter 10.

the outcome variable), and X stands for exposure to the intervention. Each row designates a different group, and time is portrayed moving from left to right. Thus, in Row 2 (a basic pretest–posttest design), the top line represents the group that was randomly assigned (R) to an intervention (X) and from which data were collected prior to (O_1) and after (O_2) the intervention. The second row is the control group, which differs from the experimental group only by absence of the treatment (no X). (Note that some information in the "drawbacks" column of Table 9.5 is not discussed until Chapter 10.)

Basic Experimental Designs

Earlier in this chapter, we described a study that tested the effect of gentle massage on pain in nursing home residents. This example illustrates a simple design that is sometimes called a **posttest-only design** (or **after-only design**) because data on the dependent variable are collected only once—after randomization and completion of the intervention.

A second basic design involves the collection of baseline data, as shown in the flow chart (Figure 9.1). Suppose we hypothesized that convective airflow blankets are more effective than conductive water-flow blankets in cooling critically ill febrile patients. Our design involves assigning patients to the two types of blankets (the independent variable) and measuring the dependent variable (body temperature) twice, before and after the intervention. This design allows us to examine whether one blanket type is more effective than the other in *reducing* fever—that is, with this design researchers can examine *change*. This design is a **pretest–posttest design** or a **before–after design**. Many pretest–posttest designs include data collection at multiple postintervention points (sometimes called *repeated measures designs*, as noted in Chapter 8). Designs that involve collected data multiple times from two groups can be described as mixed designs: analyses can examine both differences *between* groups and changes *within* groups over time.

These basic designs can be "tweaked" in various ways—for example, the design could involve comparison of three or more groups or could have a wait-listed control group. These designs are included in Table 9.5.

Example of a pretest–posttest experimental design: Wentworth and colleagues (2009) tested the efficacy of a 20-minute massage on tension, anxiety, and pain in patients awaiting invasive cardiovascular procedures. Outcomes were measured before and after the massage.

Factorial Design

Most experimental designs involve manipulating only one independent variable, but it is possible to manipulate two or more variables simultaneously. Suppose we were interested in comparing two therapies for premature infants: tactile stimulation versus auditory stimulation. We also want to learn if the daily *amount* of stimulation (15, 30, or 45 minutes) affects infants' progress. The outcomes are measures of infant development (e.g., weight gain, cardiac responsiveness). Figure 9.2 illustrates the structure of this RCT.

This **factorial design** allows us to address three research questions:

1. Does auditory stimulation have a more beneficial effect on premature infants' development than tactile stimulation, or vice versa?
2. Is the duration of stimulation (independent of type) related to infant development?
3. Is auditory stimulation most effective when linked to a certain dose and tactile stimulation most effective when coupled with a different dose?

The third question shows the strength of factorial designs: they permit us to test not only **main effects**

Type of stimulation

Daily dose	Auditory A1		Tactile A2	
15 Min. B1	A1	B1	A2	B1
30 Min. B2	A1	B2	A2	B2
45 Min. B3	A1	B3	A2	B3

FIGURE 9.2 Example of a 2 × 3 factorial design.

(effects from experimentally manipulated variables, as in questions 1 and 2), but also **interaction effects** (effects from combining treatments). It may be insufficient to say that auditory stimulation is better than tactile stimulation (or vice versa) and that 45 minutes of daily stimulation is more effective than 15 or 30 minutes. How these two variables interact (how they behave in combination) is also of interest. Our results may indicate that 45 minutes of auditory stimulation is the most beneficial treatment. We could not have learned this by conducting two separate studies that manipulated one independent variable and held the second one constant.

In factorial experiments, people are randomly assigned to a specific combination of conditions. In our example in Figure 9.2, infants would be assigned randomly to one of six **cells**—that is, six treatment conditions or boxes in the diagram. The two independent variables in a factorial design are the **factors**. Type of stimulation is factor A and amount of daily exposure is factor B. Level 1 of factor A is auditory and level 2 of factor A is tactile. When describing the dimensions of the design, researchers refer to the number of **levels**. The design in Figure 9.2 is a 2 × 3 design: two levels in factor A times three levels in factor B. Factorial experiments can be performed with multiple independent variables (factors), but designs with more than three factors are rare.

Example of a factorial design: Munro and colleagues (2009) used a 2 × 2 factorial design to test treatments to prevent ventilator-associated pneumonia in critically ill adults. Patients were randomly assigned to 1 of 4 conditions: 0.12% solution chlorhexidine oral swab twice daily, toothbrushing three times daily, both treatments, or neither treatment.

Crossover Design

Thus far, we have described RCTs in which different people are randomly assigned to different treatments. For instance, in the previous example, infants exposed to auditory stimulation were not the same infants as those exposed to tactile stimulation. A **crossover design** involves exposing the same people to more than one condition. This type of within-subjects design has the advantage of ensuring the highest possible equivalence among participants exposed to different conditions—the groups being compared are equal with respect to age, weight, health, and so on because they are composed of the same people.

Because randomization is a signature characteristic of an experiment, participants in a crossover design must be randomly assigned to different orderings of treatments. For example, if a crossover design were used to compare the effects of auditory and tactile stimulation on infant development, some infants would be randomly assigned to receive auditory stimulation first, and others would be assigned to receive tactile stimulation first. When there are three or more conditions to which participants will be exposed, the procedure of **counterbalancing** can be used to rule out ordering effects. For example, if there were three conditions (A, B, C), participants would be randomly assigned to one of six counterbalanced orderings:

A, B, C	A, C, B
B, C, A	B, A, C
C, A, B	C, B, A

Although crossover designs are extremely powerful, they are inappropriate for certain research questions because of the problem of **carry-over effects**. When people are exposed to two different treatments or conditions, they may be influenced in the second condition by their experience in the first condition. As one example, drug studies rarely use a crossover design because drug B administered *after* drug A is not necessarily the same treatment as drug B administered *before* drug A. When carry-over effects are a potential concern, researchers often have a **washout period** in between the treatments (i.e., a period of no treatment exposure).

Crossover designs usually involve treatments administered in a time sequence. Crossover designs can, however, involve simultaneous tests on two sides of a person's body.

Example of a crossover design: Pinar and colleagues (2009) tested two leg bag products (with and without latex) on a sample of men postradical prostatectomy. Each product was tested, in a randomized order, for 4 to 5 days.

Strengths and Limitations of Experiments

In this section, we explore the reasons why experimental designs are held in high esteem and examine some limitations.

Experimental Strengths

An experimental design is the gold standard for testing interventions because it yields strong evidence about intervention effects. Through randomization and the use of a comparison condition, experimenters come as close as possible to attaining the "ideal" counterfactual. Experiments offer greater corroboration than any other approach that, *if* the independent variable (e.g., diet, drug, teaching approach) is manipulated, *then* certain consequences in the dependent variable (e.g., weight loss, recovery, learning) may be expected to ensue. The great strength of RCTs, then, lies in the confidence with which causal relationships can be inferred. Through the controls imposed by manipulation, comparison, and—especially—randomization, alternative explanations can often be ruled out or discredited. It is because of these strengths that meta-analyses of RCTs, which integrate evidence from multiple studies using an experimental design, are at the pinnacle of evidence hierarchies for questions about treatment (Figure 2.1, p. 28).

Experimental Limitations

Despite the benefits of experimental research, this type of design also has limitations. First, there are often constraints that make an experimental approach impractical or impossible. These constraints are discussed later in this chapter.

TIP: Shadish and colleagues (2002) described 10 situations that are especially conducive to randomized experiments; these are summarized in a table in the Toolkit.

Experiments are sometimes criticized for their artificiality. Part of the difficulty lies in the requirements for randomization and then comparable treatment within groups, with strict adherence to protocols. In ordinary life, the way we interact with people is not random. Another aspect of experiments that is considered artificial is the focus on only a handful of variables while holding all else constant. This requirement has been criticized as being reductionist and as artificially constraining human experience. Experiments that are undertaken without a guiding theoretical framework are sometimes criticized for suggesting causal connections without any explanation for *why* the intervention affected observed outcomes.

A problem with RCTs conducted in clinical settings is that it is often clinical staff, rather than researchers, who administer an intervention; therefore, it can sometimes be difficult to determine if those in the intervention group actually received the treatment and if those in the control group did not. It may be especially difficult to maintain the integrity of the intervention and control conditions if the study period extends over time. Moreover, clinical studies are conducted in environments over which researchers may have little control—and control is a critical factor in RCTs. McGuire and colleagues (2000) have described some issues relating to the challenges of testing interventions in clinical settings.

Sometimes a problem emerges if participants have discretion about participation in the treatment. Suppose, for example, that we randomly assigned patients with HIV infection to a special support group intervention or to a control group. Experimental subjects who elect not to participate in the support groups, or who participate infrequently, actually are in a "condition" that looks more like the control condition than the experimental one. The treatment is diluted through nonparticipation, and it may become difficult to detect any treatment effects, no matter how effective it might otherwise have been. We discuss this at greater length in the next chapter.

Another potential problem is the **Hawthorne effect**, a placebo-type effect caused by people's expectations. The term is derived from a set of experiments conducted at the Hawthorne plant of the Western Electric Corporation in which various environmental conditions, such as light and working hours, were varied to test their effects on worker productivity. Regardless of what change was introduced, that is, whether the light was made

better or worse, productivity increased. Knowledge of being included in the study (not just knowledge of being in a particular group) appears to have affected people's behavior, thus obscuring the effect of the treatment.

In sum, despite the superiority of RCTs for testing causal hypotheses, they are subject to a number of limitations, some of which may make them difficult to apply to real-world problems. Nevertheless, with the growing demand for evidence-based practice, true experimental designs are increasingly being used to test the effects of nursing interventions.

QUASI-EXPERIMENTS

Quasi-experiments, called *controlled trials without randomization* in the medical literature, involve an intervention but they lack randomization, the signature of a true experiment. Some quasi-experiments even lack a control group. The signature of a quasi-experimental design, then, is an intervention in the absence of randomization.

Quasi-Experimental Designs

The most widely used quasi-experimental designs are summarized in Table 9.6, which depicts designs using the schematic notation we introduced earlier.

Nonequivalent Control Group Designs

The **nonequivalent control group pretest–posttest design** involves two groups of participants, from whom outcome data are collected before and after implementing an intervention. For example, suppose we wished to study the effect of a new hospital-wide model of care that involved having a patient care facilitator (PCF) be the primary point person for all patients during their stay. Our main outcome is patient satisfaction. The new system is being implemented throughout the hospital, and so, randomization is not possible. For comparative purposes, we decide to collect data in a similar hospital that is not instituting the PCF model. Data on patient satisfaction is collected in both hospitals at baseline, before the change is made, and again after its implementation.

The first row of Table 9.6 depicts this study symbolically. The top line represents the experimental (PCF) hospital, and the second row is the comparison hospital. This diagram is identical to the experimental pretest–posttest design (see Table 9.5), *except* there is no "R"—participants have not been randomized to groups. The design in Table 9.6 is weaker because *it cannot be assumed that the experimental and comparison groups are equivalent at the outset.* Because there is no randomization, quasi-experimental comparisons are farther from an ideal counterfactual than experimental comparisons. The design is nevertheless strong, because baseline data allow us to assess whether patients in the two hospitals had similar satisfaction initially. If the comparison and experimental groups are similar at baseline, we could be relatively confident inferring that any posttest difference in satisfaction was the result of the new care model. If patient satisfaction is different initially, however, it will be difficult to interpret posttest differences. Note that in quasi-experiments, the term **comparison group** is often used in lieu of *control group* to refer to the group against which treatment group outcomes are evaluated.

Now, suppose we had been unable to collect baseline data. This design, diagramed in Row 2 of Table 9.6, has a major flaw. We no longer have information about the initial equivalence of the two hospitals. If we find that patient satisfaction in the experimental hospital is higher than that in the control hospital at posttest, can we conclude that the new care delivery method *caused* improved satisfaction? An alternative explanation for posttest differences is that patient satisfaction in the two hospitals differed initially. Campbell and Stanley (1963) called this *nonequivalent control group posttest-only design* preexperimental rather than quasi-experimental because of its fundamental weakness—although Shadish, and colleagues (2002), in their more recent book on causal inference, simply called this a weaker quasi-experimental design.

TABLE 9.6 Design Alternatives: Selected Quasi-Experimental Designs

TYPE OF DESIGN	SCHEMATIC DIAGRAM	SITUATIONS THAT ARE BEST SUITED	DRAWBACKS
1. Nonequivalent control group, pretest–posttest design	O_1 X O_2 O_1 O_2	Attractive when an entire unit must get the intervention and a similar unit not getting the intervention is available	a. Selection threat* remains a nearly intractable problem, but less so than when there is no pretest b. History threat* also a possibility
2. Nonequivalent control group, posttest only design	X O_1 O_1	A reasonable choice only when there is some a priori knowledge about comparability of groups with regard to key outcomes	Extremely vulnerable to selection threat,* possibility of other threats as well, especially history threat*
3. One-group pretest–posttest design	O_1 X O_2	A reasonable choice only when intervention impact is expected to be dramatic and other potential causes have little credibility	Typically provides very weak support for causal inference—vulnerable to many internal validity threats (maturation, history, etc.)*
4. Time series design	O_1 O_2 O_3 O_4 X O_5 O_6 O_7 O_8	a. Good option when there are abundant data on key outcome in existing records b. Addresses maturation threat and change from secular trends & random fluctuation	a. Complex statistical analysis that is most appropriate with very large number of data points (100+) b. History threat* remains, and (sometimes) selection threat* if the population changes over time
5. Time series nonequivalent control group design	O_1 O_2 O_3 O_4 X O_5 O_6 O_7 O_8 O_1 O_2 O_3 O_4 O_5 O_6 O_7 O_8	Attractive when an entire unit/institution adopts the intervention and a similar unit not adopting it is available, *and* if comparable data are readily available in records of both	a. Selection threat* remains, as two units or institutions are rarely identical b. Analyses may be very complex
6. Time series with withdrawn and reinstituted treatment	O_1 O_2 X O_3 O_4 –X O_5 O_6 X O_7 O_8	*Attractive if effects of an intervention are short-term	a. May be untenable to assume that there are no carryover effects b. May be difficult ethically to withdraw treatment if it is efficacious

KEY: X = Intervention
O = Observation or measurement of the dependent variable/outcome

*Validity threats are discussed in Chapter 10.

Example of a nonequivalent control group pretest–posttest design: Yuan and colleagues (2009) tested the effectiveness of an exercise intervention on nurses' physical fitness. The researchers used nurses from different units of a medical center in Taiwan to be in either an intervention group or a comparison group.

Sometimes researchers use *matching* within a pretest–posttest nonequivalent control group design to ensure that the groups are, in fact, equivalent on at least some key variables related to the outcomes. For example, if an intervention was designed to reduce patient anxiety, then it might be desirable to not only *measure* preintervention anxiety in the intervention and comparison group, but to take steps to ensure that the groups' anxiety levels were comparable by matching participants' initial anxiety. Because matching on more than a couple variables is unwieldy, a more sophisticated method of matching, called **propensity matching**, can be used by researchers with statistical sophistication. This method involves the creation of a single **propensity score** that captures the conditional probability of exposure to a treatment given various preintervention characteristics. Experimental and comparison group members can then be matched on this score (Qin et al., 2008). Both conventional and propensity matching are most easily implemented when there is a large pool of potential comparison group participants from which good matches to treatment group members can be selected.

In lieu of using a contemporaneous nonrandomized comparison group, researchers sometimes use a **historical comparison group**. That is, comparison data are gathered about a group of people before implementing the intervention. Even when the people are from the same institutional setting, however, it is risky to assume that the two groups are comparable, or that the environments are comparable in all respects except for the new intervention. There remains the possibility that something other than the intervention could account for any observed differences in outcomes.

Example of a historical comparison group: Swadener-Culpepper and colleagues (2008) studied the effect of continuous lateral rotation therapy on patients at high risk for pulmonary complications. Length of stay for those receiving the therapy was compared to that for a high-risk historical comparison group.

Time Series Designs

In the designs just described, a control group was used but randomization was not, but some quasi-experiments have neither. Suppose that a hospital implemented rapid response teams (RRTs) in its acute care units. Administrators want to examine the effects on patient outcomes (e.g., unplanned admissions to the ICU, mortality rate) and nurse outcomes (e.g., stress). For the purposes of this example, assume no other hospital could serve as a good comparison. The only kind of comparison that can be made is a before—after contrast. If RRTs were implemented in January, one could compare the mortality rate (for example) during the 3 months before RRTs with the mortality rate during the subsequent 3-month period. The schematic representation of such a study is shown in the third row of Table 9.6.

This one-group pretest–posttest design seems straightforward, but it has weaknesses. What if either of the 3-month periods is atypical, apart from the innovation? What about the effects of any other policy changes inaugurated during the same period? What about the effects of external factors that influence mortality, such as a flu outbreak or seasonal migration? This design (also called preexperimental by Campbell and Stanley) cannot control these factors.

TIP: One-group pretest–posttest designs are not always unproductive. For example, if a study tested a brief teaching intervention, with baseline knowledge data obtained immediately before the intervention and posttest knowledge data collected immediately after it, it may be reasonable to infer that the intervention is the most plausible explanation for knowledge gains.

In our RRT example, the design could be modified so that some alternative explanations for

changes in mortality could be ruled out. One such design is the **time series design** (sometimes called an *interrupted time series design*), diagramed in Row 4 of Table 9.6. In a time series design, data are collected over an extended period and an intervention is introduced during that period. In the diagram, O_1 through O_4 represent four separate instances of data collection on an outcome before treatment, X is the introduction of the intervention, and O_5 through O_8 represent four posttreatment observations. In our example, O_1 might be the number of deaths in January through March in the year before the new RRT system, O_2 the number of deaths in April through June, and so forth. After RRTs are implemented, data on mortality are similarly collected for four consecutive 3-month periods, giving us observations O_5 through O_8.

Even though the time series design does not eliminate all problems of interpreting changes in mortality, the extended time period strengthens the ability to attribute change to the intervention. Figure 9.3 demonstrates why this is so. The two line graphs (*A* and *B*) in the figure show two possible outcome patterns for eight mortality observations. The vertical dotted line in the center represents the timing of the RRT system. Patterns *A* and *B* both reflect a feature common to most time series studies—fluctuation from one data point to another. These fluctuations are normal. One would not expect that, if 480 patients died in a hospital in 1 year, the deaths would be spaced evenly with 40 per month. It is precisely because of these fluctuations that the one-group pretest–posttest design, with only one observation before and after the intervention, is so weak.

Let us compare the interpretations that can be made for the outcomes shown in Figure 9.3. In both patterns *A* and *B*, mortality decreased between O_4 and O_5, immediately after RRTs were implemented. In *B*, however, mortality rose at O_6 and continued to rise at O_7. The decrease at O_5 looks similar to other apparently haphazard fluctuations in mortality. In *A*, by contrast, the number of deaths decreases at O_5 and remains relatively low for subsequent observations. There may be other explanations for a change in the mortality rate, but the time series design does permit us to rule out the possibility that the data reflect unstable measurements at only two points in time. If we had used a simple pretest–posttest design, it would have been analogous to obtaining the measurements at O_4 and

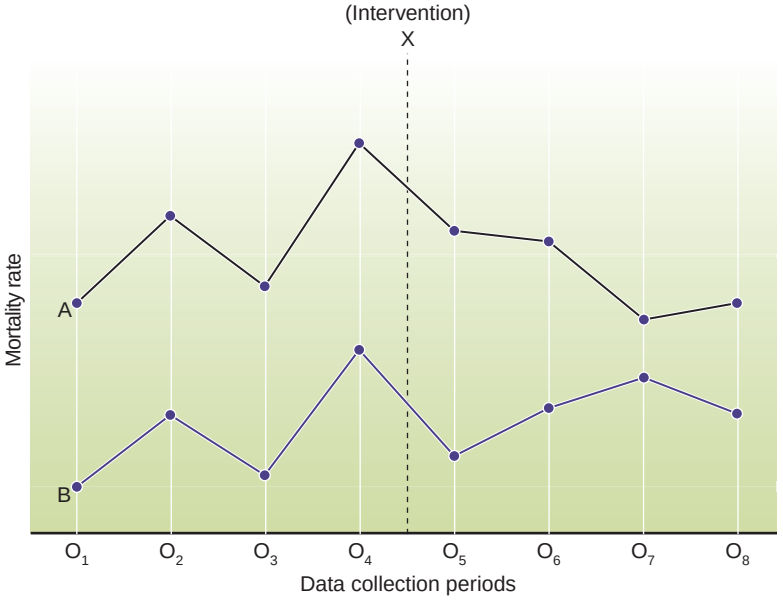


FIGURE 9.3 Two possible time series outcome patterns.

O_5 of Figure 9.3 only. The outcomes in both *A* and *B* are the same at these two time points. The broader time perspective leads us to draw different conclusions about the effects of RRTs. Nevertheless, the absence of a comparison group means that the design is far from yielding an ideal counterfactual.

Time series designs are often especially important in *quality improvement studies*, because in such efforts randomization is rarely possible, and only one institution is involved in the inquiry.

Example of a time series design: Kratz (2008) used a time series design to test the effects of implementing research-based protocols to decrease negative outcomes associated with delirium and acute confusion. Kratz used 3 years of hospital records data prior to and 4 years of records data after implementing the new protocols, for such outcomes as patient falls and use of restraints.

One drawback of a time series design is that a large number of data points—100 or more—is recommended for a traditional analysis (Shadish et al., 2002), and the analyses are complex. Nurse researchers are, however, beginning to use a little-known but versatile and compelling approach called *statistical process control* to assess effects when they have collected data sequentially over a period of time before and after implementing an intervention or practice change (Polit & Chaboyer, in review).

A powerful quasi-experimental design results when time series and nonequivalent control group designs are combined (Row 5 of Table 9.6). In the example just described, a time series nonequivalent control group design would involve collecting data over an extended period from both the hospital introducing the RRTs and another similar hospital not implementing RRTs. Information from another hospital with similar characteristics would make inferences regarding the effects of RRTs more convincing because other factors influencing the trends would likely be comparable in both groups.

Numerous variations on the time series design are possible. For example, additional evidence regarding the effects of a treatment can be achieved by instituting the treatment at several different points in time, strengthening the treatment over time, or instituting the treatment at one point and then withdrawing it at a later point, sometimes with reinstitution (Row 6 of Table 9.6). Clinical nurse researchers may be in a good position to use such time series designs because many measures of patient functioning are routinely made at multiple points over an extended period.

Example of a time series design with withdrawal and reinstitution: Hicks-Moore (2005) studied the effect of relaxing music at mealtime on agitated behaviors of nursing home residents with dementia. Music was introduced in week 2, removed in week 3, and then reinstituted in week 4. The pattern of agitated behaviors was consistent with the hypothesis that relaxing music has a calming effect.

A particular application of a time series approach is called **single-subject experiments** (*N-of-1 studies*). Single-subject studies use time series designs to gather information about intervention effects based on a single patient (or a small number of patients) under controlled conditions. The most basic single-subject design involves a baseline phase of data gathering (A) and an intervention phase (B), yielding an **AB design**. If the treatment is withdrawn, it would be an **ABA design**; if a withdrawn treatment is reinstituted, it would be an **ABAB design**. Portney and Watkins (2000) offer valuable guidance about single-subject studies in clinical settings.

Example of a single-subject ABAB design: Elliott and Horgas (2009) used an ABAB design in which the intervention (a scheduled dose of acetaminophen) was administered, withdrawn, and then reinstituted in three people with dementia. Data on pain-related behaviors were collected daily for 24 days.

Other Quasi-Experimental Designs

Several other quasi-experimental designs offer alternatives to RCTs. One such design, the **regression discontinuity design**, will not be elaborated on here because it is rarely used in nursing studies. This design, which involves *systematic* assignment of people to groups based on cut-off scores on a preintervention measure (e.g., giving an intervention to the most severely ill patients), is considered attractive from an ethical standpoint and merits

consideration. Its features have been described in the nursing literature by Atwood and Taylor (1991).

Earlier in this chapter, we described partially randomized patient preference or PRPP. This design has advantages in terms of participant recruitment to participate in a study, because those with a strong preference get to choose their treatment condition. Those without a strong preference are randomized, but those *with* a preference are given the condition they prefer and are followed up as part of the study. The two randomized groups are part of a true experiment, but the two groups who get their preference are in a quasi-experiment. This design can yield valuable information about the kind of people who prefer one condition over another. The evidence of treatment effectiveness is weak in the quasi-experimental segment because the people who elected a certain treatment likely differ from those who opted for the alternative—and these preintervention differences, rather than the alternative treatments, could account for any observed differences in outcomes. Yet, evidence from the quasi-experiment could usefully support or qualify evidence from the experimental portion of the study.

Example of a PRPP design: Coward (2002) used a PRPP design in a pilot study of a support group intervention for women with breast cancer. She found that the majority of women did *not* want to be randomized, but rather had a strong preference for either being in or not being in the support group. Her article describes the challenges she faced.

Another quasi-experimental approach—often embedded within a true experiment—is a **dose-response** design in which the outcomes of those receiving different doses of a treatment—not as a result of randomization—are compared. For example, in complex and lengthy interventions, some people attend more sessions or get more intensive treatment than others. The rationale for a quasi-experimental dose-response analysis is that if a larger dose corresponds to better outcomes, this provides supporting evidence for inferring that the treatment caused the outcome. The difficulty, however, is that people tend to get different doses of the treatment because of differences in motivation, physical function, or other characteristics that could be driving outcome differences—and not the different doses themselves. Nevertheless, when a dose-response analyses may yield useful information.

Example of a dose-response analysis within a true experiment: Lai and Good (2005) randomly assigned community dwelling elders who had difficulty sleeping to a control group or to an intervention group that listened to 45-minute sedative music tapes at bedtime. Those in the intervention group experienced significantly better sleep quality than those in the control group. Moreover, over the 3-week study period, sleep improved weekly, which suggested a cumulative dose effect.

Experimental and Comparison Conditions

Researchers using a quasi-experimental approach, like those adopting an experimental design, should strive to develop strong interventions that provide an opportunity for a fair test, and should develop protocols documenting what the interventions entail. Researchers need to be especially careful in understanding and documenting the counterfactual in quasi-experiments. In the case of nonequivalent control group designs, this means understanding the conditions to which the comparison group is exposed. In our example of using a hospital with traditional nursing systems as a comparison for the new primary nursing system, the nature of that traditional system should be fully understood. In time series designs, the counterfactual is the condition existing before implementing the intervention. Blinding should be used, to the extent possible—indeed, this is often more feasible in a quasi-experiment than in an RCT.

Strengths and Limitations of Quasi-Experiments

A major strength of quasi-experiments is that they are practical. In clinical settings, it is often impossible to conduct true experimental tests of nursing interventions. Quasi-experimental designs introduce some research control when full experimental rigor is not possible.

Another advantage of quasi-experiments is that patients are not always willing to relinquish control over their treatment condition. Indeed, there is some evidence that people are increasingly unwilling to volunteer to be randomized in clinical trials (Gross & Fogg, 2001). Quasi-experimental designs, because they do not involve random assignment, are likely to be acceptable to a broader group of people. This, in turn, has implications for the generalizability of the results—but the problem is that the results may be less conclusive.

Thus, researchers using quasi-experimental designs need to be cognizant of their weaknesses and need to take steps to counteract those weaknesses or at least take them into account in interpreting results. When a quasi-experimental design is used, there may be several **rival hypotheses** competing with the experimental manipulation as explanations for the results. (This issue relates to *internal validity* and is discussed further in Chapter 10.) Take as an example the case in which we administer a special diet to frail nursing home residents to assess its effects on weight gain. If we use no comparison group or if we use a nonequivalent control group and then observe a weight gain, we must ask the questions: Is it *plausible* that some other factor caused the gain? Is it *plausible* that pretreatment differences between the experimental and comparison groups resulted in differential gain? Is it *plausible* that the elders, on average, gained weight simply because the most frail died or were transferred to a hospital? If the answer is "yes" to any of these questions, then inferences about the causal effect of the intervention are weakened. The plausibility of any particular rival explanation cannot be answered unequivocally. Usually, judgment must be exercised. Because the conclusions from quasi-experiments ultimately depend in part on human judgment, rather than on more objective criteria, cause-and-effect inferences are less compelling.

NONEXPERIMENTAL RESEARCH

Many research questions—including ones seeking to establish causal relationships—cannot be addressed with an experimental or quasi-experimental design. For example, at the beginning of this chapter, we posed this prognosis question: Do birth weights under 1,500 grams *cause* developmental delays in children? Clearly, we cannot manipulate birth weight, the independent variable. Babies are born with weights that are neither random nor subject to research control. One way to answer this question is to compare two groups of infants—babies with birth weights above and below 1,500 grams at birth—in terms of their subsequent development. When researchers do not intervene by manipulating the independent variable, the study is **nonexperimental**, or, in the medical literature, **observational**.

Most nursing studies are nonexperimental, mainly because most human characteristics (e.g., birth weight, ethnicity, lactose intolerance) cannot be experimentally manipulated. Also, many variables that could *technically* be manipulated cannot be manipulated ethically. For example, if we were studying the effect of prenatal care on infant mortality, it would be unethical to provide such care to one group of pregnant women while deliberately depriving a randomly assigned second group. We would need to locate a naturally occurring group of pregnant women who had not received prenatal care. Their birth outcomes could then be compared with those of women who had received appropriate care. The problem, however, is that the two groups of women are likely to differ in terms of many other characteristics, such as age, education, and income, any of which individually or in combination could affect infant mortality, independent of prenatal care. This is precisely why experimental designs are so strong in demonstrating cause-and-effect relationships. Many nonexperimental studies are designed to explore causal relationships when experimental work is not possible—although, some studies have primarily a descriptive intent.

Correlational Cause-Probing Research

When researchers study the effect of a potential *cause* that they cannot manipulate, they use **correlational designs** to examine relationships between

variables. A **correlation** is a relationship or association between two variables, that is, a tendency for variation in one variable to be related to variation in another. For example, in human adults, height and weight are correlated because there is a tendency for taller people to weigh more than shorter people.

As mentioned early in this chapter, one criterion for causality is that an empirical relationship (correlation) between variables must be demonstrated. It is risky, however, to infer causal relationships in correlational research. In experiments, researchers have direct control over the independent variable; the experimental treatment can be administered to some and withheld from others, and the two groups can be equalized with respect to everything except the independent variable through randomization. In correlational research, on the other hand, investigators do not control the independent variable, which often has already occurred. Groups being compared could differ in many respects that could affect outcomes of interest. Although correlational studies are inherently weaker than experimental studies in elucidating cause-and-effect relationships, different designs offer different degrees of supportive evidence.

Retrospective Designs

Studies with a **retrospective design** are ones in which a phenomenon existing in the present is linked to phenomena that occurred in the past. The signature of a retrospective study is that the researcher begins with the dependent variable (the effect) and then examines whether it is correlated with one or more previously occurring independent variables (potential causes).

Most early studies of the smoking–lung cancer link used a retrospective **case-control design**, in which researchers began with a group of people who had lung cancer (*cases*) and another group who did not (*controls*). The researchers then looked for differences between the two groups in antecedent behaviors or conditions, such as smoking.

In designing a case-control study, researchers try to identify controls without the disease or condition who are as similar as possible to the cases with regard to key confounding variables (e.g., age, gender). Researchers sometimes use matching or other techniques to control for confounding variables. (Sometimes they opt to match two or more controls for each case). To the degree that researchers can demonstrate comparability between cases and controls with regard to confounding traits, inferences regarding the presumed cause of the disease are enhanced. The difficulty, however, is that the two groups are almost never totally comparable with respect to all potential factors influencing the dependent variable.

Example of a case-control design: Swenson and colleagues (2009) used a case-control design to assess risk factors for lymphedema following breast cancer surgery. Women with and without lymphedema were matched on type of axillary surgery and surgery date, and then compared to such antecedent risk factors as weight, number of positive nodes, and treatments received.

Not all retrospective studies can be described as using a case-control design. Sometimes researchers use a retrospective approach to identify risk factors for different *amounts* of a problem or condition. That is, the outcome is not "caseness" but rather *degree* of some condition. For example, a retrospective design might be used to identify factors predictive of the length of time new mothers breastfed their infants. Essentially, such a design is intended to understand factors that *cause* women to make different breastfeeding decisions.

Retrospective studies are often cross-sectional, with data on both the dependent and independent variables collected at a single point in time. In such studies, data for the independent variable are based on recollection (retrospection). One problem, however, is that recollection is often less accurate than contemporaneous measurement. Asking people if they had a headache at any time in the previous 12 months might not be difficult to answer, but asking them to report how many times they had a headache, or what it felt like to have a headache 6 months ago, is likely to result in unreliable answers.

Example of a retrospective design: Musil and colleagues (2009) used cross-sectional data in their retrospective study designed to identify antecedent factors to predict depressive symptoms in grandmothers raising their grandchildren. The independent

variables included family stresses and strains, social support, and demographic variables such as age and employment status.

Prospective Nonexperimental Designs

In correlational studies with a **prospective design** (called a **cohort design** in medical circles), researchers start with a presumed cause and then go forward in time to the presumed effect. For example, we might want to test the hypothesis that rubella during pregnancy (the independent variable) is related to birth defects (the dependent variable). To test this hypothesis prospectively, we would begin with a sample of pregnant women, including some who contracted rubella during pregnancy and others who did not. The subsequent occurrence of congenital anomalies would be assessed for all participants, and we would examine whether women with rubella were more likely than other women to bear infants with birth defects.

Prospective studies are more costly than retrospective studies, in part because prospective studies require at least two rounds of data collection. A substantial follow-up period may be needed before the outcome of interest occurs, as is the case in prospective studies of cigarette smoking and lung cancer. Also, prospective designs require large samples if the outcome of interest is rare, as in the example of malformations associated with maternal rubella. Another issue is that in a good prospective study, researchers take steps to confirm that all participants are free from the effect (e.g., the disease) at the time the independent variable is measured, and this may be difficult or expensive to do. For example, in prospective smoking–lung cancer studies, lung cancer may be present initially but not yet diagnosed.

Despite these issues, prospective studies are considerably stronger than retrospective studies. In particular, any ambiguity about whether the presumed cause occurred before the effect is resolved in prospective research if the researcher has confirmed the initial absence of the effect. In addition, samples are more likely to be representative, and investigators may be in a position to impose controls to rule out competing explanations for the results.

TIP: The term "prospective" is not synonymous with "longitudinal." Although most nonexperimental prospective studies *are* longitudinal, prospective studies are not *necessarily* longitudinal. Prospective means that information about a possible cause is obtained prior to information about an effect. RCTs are inherently prospective because the researcher introduces the intervention and then determines its effect. An RCT that collected data 1 hour after an intervention would be prospective, but not longitudinal.

Some prospective studies are exploratory. Researchers sometimes measure a wide range of possible "causes" at one point in time, and then examine an outcome of interest at a later point (e.g., length of stay in hospital). Such studies are usually stronger than retrospective studies if it can be determined that the outcome was not present initially because time sequences are clear. They are not, however, as powerful as prospective studies that involve specific *a priori* hypotheses and the comparison of cohorts known to differ on a presumed cause. Researchers doing exploratory retrospective or prospective studies are sometimes accused of going on "fishing expeditions" that can lead to erroneous conclusions because of spurious or idiosyncratic relationships in a particular sample of participants.

Example of a prospective nonexperimental study: Wiklund and colleagues (2009) conducted a prospective cohort study of first-time mothers to examine the effect of mode of delivery (vaginal versus cesarean) on changes in the mothers' personality from predelivery to 9 months after delivery.

Natural Experiments

Researchers are sometimes able to study the outcomes of a "**natural experiment**" in which a group exposed to a phenomenon with potential health consequences is compared with a nonexposed group. Natural experiments are nonexperimental because the researcher does not intervene, but they are called "natural *experiments*" if people are affected essentially at random. For example, the psychological well-being of people living in a community struck with a natural disaster (e.g., a volcanic eruption) could be compared with the well-being of people living in a similar but unaffected community to

determine the toll exacted by the disaster (the independent variable). Note that the independent variable or "cause" does not need to be a "natural" phenomenon. It could, for example, be a fire or winning the lottery. Moreover, the groups being compared do not need to be different people; if preevent measures have been obtained, before–after comparisons might be profitable.

Example of a natural experiment: Liehr and colleagues (2004) were in the midst of collecting data from healthy students over a 3-day period (September 10 to 12, 2001) when the events of September 11 unfolded. The researchers seized the opportunity to examine what people go through in the midst of stressful upheaval. Both pre- and post-tragedy data were available for the students' blood pressure, heart rate, and television viewing.

Path Analytic Studies

Researchers interested in testing theories of causation based on nonexperimental data are increasingly using a technique known as **path analysis** (or similar techniques). Using sophisticated statistical procedures, researchers test a hypothesized causal chain among a set of independent variables, mediating variables, and a dependent variable. Path analytic procedures, described briefly in Chapter 18, allow researchers to test whether nonexperimental data conform sufficiently to the underlying model to justify causal inferences. Path analytic studies can be done within the context of both cross-sectional and longitudinal designs, the latter providing a stronger basis for causal inferences because of the ability to sort out time sequences.

Example of a path analytic study: Chen and Tzeng (2009) tested a model to explain adherence to pelvic floor muscle exercise among women with urinary incontinence. Their path analysis tested hypothesized causal pathways between adherence on the one hand and self-efficacy, exercise knowledge and attitudes, and severity of urine loss on the other.

Descriptive Research

A second broad class of nonexperimental studies is **descriptive research**. The purpose of descriptive studies is to observe, describe, and document aspects of a situation as it naturally occurs and sometimes to serve as a starting point for hypothesis generation or theory development.

Descriptive Correlational Studies

Sometimes researchers are better able to simply describe relationships than to comprehend causal pathways. Many research problems are cast in noncausal terms. We ask, for example, whether men are less likely than women to bond with their newborn infants, not whether a particular configuration of sex chromosomes *caused* differences in parental attachment. Unlike other types of correlational research—such as the cigarette smoking and lung cancer investigations—the aim of **descriptive correlational research** is to describe relationships among variables rather than to support inferences of causality.

Example of a descriptive correlational study: Jacob and colleagues (2010) conducted a descriptive correlational study to examine the relationship between respiratory symptoms and pain experiences in children and adolescents with sickle cell disease.

Studies designed to address diagnosis/assessment questions—that is, whether a tool or procedure yields accurate assessment or diagnostic information about a condition or outcome—typically involve descriptive correlational designs. Procedures are discussed in Chapter 15.

Univariate Descriptive Studies

The aim of some descriptive studies is to describe the frequency of occurrence of a behavior or condition, rather than to study relationships. **Univariate descriptive studies** are not necessarily focused on only one variable. For example, a researcher might be interested in women's experiences during menopause. The study might describe the frequency of various symptoms, the average age at menopause, and the percentage of women using medications to alleviate symptoms. The study involves multiple variables, but the primary purpose is to describe the status of each and not to relate them to one another.

Two types of descriptive study come from the field of epidemiology. **Prevalence studies** are done to estimate the prevalence rate of some condition (e.g., a disease or a behavior, such as smoking) at a particular point in time. Prevalence studies rely on

cross-sectional designs in which data are obtained from the population at risk of the condition. The researcher takes a "snapshot" of the population at risk to determine the extent to which the condition of interest is present. The formula for a **prevalence rate** (PR) is:

$$\frac{\text{Number of cases with the condition or disease at a given point in time}}{\text{Number in the population at risk of being a case}} \times K$$

K is the number of people for whom we want to have the rate established (e.g., per 100 or per 1,000 population). When data are obtained from a sample (as would usually be the case), the denominator is the size of the sample, and the numerator is the number of cases with the condition, as identified in the study. If we sampled 500 adults aged 21 years and older living in a community, administered a measure of depression, and found that 80 people met the criteria for clinical depression, then the estimated prevalence rate of clinical depression would be 16 per 100 adults in that community.

Incidence studies estimate the frequency of developing *new* cases. Longitudinal designs are needed to estimate incidence because the researcher must first establish who is at risk of becoming a new case—that is, who is free of the condition at the outset. The formula for an **incidence rate** (IR) is:

$$\frac{\text{Number of new cases with the condition or disease over a given time period}}{\text{Number in the population at risk of being a case (free of the condition at the outset)}} \times K$$

Continuing with our previous example, suppose in October 2010, we found that 80 in a sample of 500 people were clinically depressed (PR = 16 per 100). To determine the 1-year incidence rate, we would reassess the sample in October 2011. Suppose that, of the 420 previously deemed *not* to be clinically depressed in 2010, 21 were now found to meet the criteria for depression. In this case, the estimated 1-year incidence rate would be 5 per 100 ((21 ÷ 420) × 100 = 5).

Prevalence and incidence rates can be calculated for subgroups of the population (e.g., for men versus women). When this is done, it is possible to calculate another important descriptive index. **Relative risk** is an estimated risk of "caseness" in one group compared with another. Relative risk is computed by dividing the rate for one group by the rate for another. Suppose we found that the 1-year incidence rate for depression was 6 per 100 women and 4 per 100 men. Women's relative risk for developing depression over the 1-year period would be 1.5, that is, women would be estimated to be 1.5 times more likely to develop depression than men. Relative risk is an important index in assessing the contribution of risk factors to a disease or condition (e.g., by comparing the relative risk for lung cancer for smokers versus nonsmokers).

Example of an incidence and prevalence study: Johansson and colleagues (2009) collected cross-sectional data to estimate the prevalence of malnutrition risk among community-dwelling older people in a Swedish municipality (14.5%). Longitudinal data were also collected to estimate the 1-year incidence rate (7.6%).

TIP: The quality of correlational studies that test hypothesized causal relationships is heavily dependent on design decisions—that is, how researchers design their studies to rule out competing causal explanations for the outcomes. Methods of enhancing the rigor of such studies are described in the next chapter. The quality of descriptive studies, by contrast, is more heavily dependent on having a good (representative) sample (Chapter 12) and high-quality measuring instruments (Chapter 14) than on design.

Strengths and Limitations of Correlational Research

The quality of a study is not necessarily related to its approach; there are many excellent nonexperimental studies as well as flawed experiments. Nevertheless, nonexperimental correlational studies have several drawbacks.

Limitations of Correlational Research

Relative to experimental and quasi-experimental research, nonexperimental studies are weak in their

ability to support causal inferences. In correlational studies, researchers work with preexisting groups that were not formed at random, but rather through **self-selection** (also known as *selection bias*). A researcher doing a correlational study cannot assume that groups being compared are similar before the occurrence of the independent variable—the hypothesized cause. Preexisting differences may be a plausible alternative explanation for any group differences on the outcome variable.

The difficulty of interpreting correlational findings stems from the fact that, in the real world, behaviors, attitudes, and characteristics are interrelated (correlated) in complex ways. An example may help to clarify the problem. Suppose we conducted a cross-sectional study that examined the relationship between level of depression in cancer patients and their social support (i.e., assistance and emotional support from others). We hypothesize that social support (the independent variable) affects levels of depression (the dependent variable). Suppose we find that the patients with weak social support are significantly more depressed than patients with strong support. We could interpret this finding to mean that patients' emotional state is influenced by the adequacy of their social supports. This relationship is diagrammed in Figure 9.4A. Yet, there are alternative explanations. Perhaps a third variable influences *both* social support and depression, such as the patients' marital status. It may be that having a spouse is a powerful influence on how depressed cancer patients feel *and* on the quality of their social support. This set of relationships is diagramed in Figure 9.4B. In this scenario, social support and depression are correlated simply because marital status affects both. A third possibility is reversed causality (Figure 9.4C). Depressed cancer patients may find it more difficult to elicit needed support from others than patients who are more cheerful or amiable. In this interpretation, the person's depression causes the amount of received social support and not the other way around. Thus, interpretations of most correlational results should be considered tentative, particularly if the research has no theoretical basis and if the design is cross-sectional.

Strengths of Correlational Research

Earlier, we discussed constraints that limit the possibility of applying experimental designs to many research problems. Correlational research will continue to play a crucial role in nursing research precisely because many interesting problems are not amenable to experimentation.

Despite our emphasis on causal inferences, it has already been noted that descriptive correlational research does not focus on understanding causal relationships. Furthermore, if the study is testing a causal hypothesis that has been deduced from an established theory, causal inferences may be possible, especially if strong designs (e.g., a prospective design) are used.

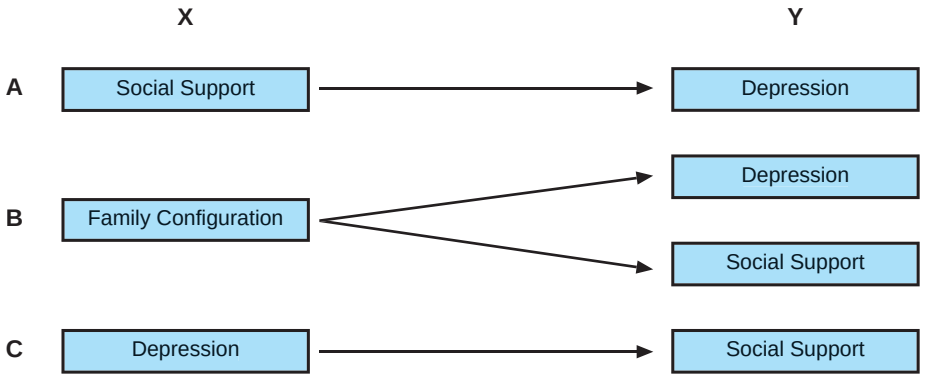


FIGURE 9.4 Alternative explanations for relationship between depression and social support in cancer patients.

Correlational research is often efficient in collecting a large amount of data about a problem. For example, it would be possible to collect extensive information about the health histories and eating habits of a large number of individuals. Researchers could then examine which health problems were associated with which diets, and could thus discover a large number of interrelationships in a relatively short amount of time. By contrast, an experimenter looks at only a few variables at a time. One experiment might manipulate foods high in cholesterol, whereas another might manipulate protein, for example.

Finally, correlational research is often strong in realism. Unlike many experimental studies, correlational research is seldom criticized for its artificiality.

TIP: It is often a good idea to design a study with as many relevant comparisons as possible. Two-group nonequivalent control group posttest-only designs are weak in part because the comparative information they yield is limited. In nonexperimental studies, multiple comparison groups can be effective in dealing with self-selection, especially if comparison groups are chosen to address competing biases. For example, in case–control studies of potential causes of lung cancer, cases would be people with lung cancer, one comparison group could comprise people with a different lung disease and a second could comprise those with no lung disorder.

DESIGNS AND RESEARCH EVIDENCE

Evidence for nursing practice depends on descriptive, correlational, and experimental research. There is often a logical progression to knowledge expansion that begins with rich description, including description from qualitative research. Descriptive studies are valuable in documenting the prevalence, nature, and intensity of health-related conditions and behaviors and are critical in the development of effective interventions. Moreover, in-depth qualitative research may suggest causal links that could be the focus of controlled quantitative research. For example, Colón-Emeric and colleagues (2006) did case studies in two nursing homes. They looked at site differences in communication patterns among the medical and nursing staff in relation to differences in information flow. Their findings suggested that a "chain of command" type communication style may limit healthcare providers' ability to provide high-quality care. The study suggests a causal hypothesis that merits greater scrutiny with a larger number of nursing homes under more controlled conditions—and also suggests possibilities for interventions. Thus, although qualitative studies are low on the standard evidence hierarchy for *confirming* causal connections (Figure 2.1), they nevertheless serve an important function.

Correlational studies also play a role in developing an evidence base for causal inferences. Retrospective case-control studies may pave the way for more rigorous (but more expensive) prospective studies. As the evidence base builds, conceptual models may be developed and tested using path analytic designs and other theory-testing strategies. These studies can provide hints about how to structure an intervention, who can most profit from it, and when it can best be instituted. Thus, nonexperimental studies can sometimes lead to innovative interventions that can be tested using experimental and quasi-experimental designs.

Many important research questions will never be answered using information from Level I (meta-analyses of RCTs) or Level II studies (RCTs) on the standard evidence hierarchy. An important example is the question of whether smoking causes lung cancer. Despite the inability to randomize people to smoking and nonsmoking groups, few people doubt that this causal connection exists. Thinking about the criteria for causality discussed early in this chapter, there is ample evidence that smoking cigarettes is correlated with lung cancer and, through prospective studies, that smoking precedes lung cancer. The large number of studies conducted has allowed researchers to control for, and thus rule out, other possible "causes" of lung cancer. There has been a great deal of consistency and coherence in the findings. And, the criterion of biologic plausibility has been met through basic physiologic research.

Thus, it may be best to think of alternative evidence hierarchies for questions relating to causality. For "therapy" questions (Table 2.1), experimental designs are the "gold standard." On the next rung of

the hierarchy for therapy questions are strong quasi-experimental designs, such as nonequivalent control group pretest–posttest designs. Further down the hierarchy are weaker quasi-experimental designs and then correlational studies.

TIP: Studies have shown that evidence from RCTs, quasi-experimental, and observational studies often do not yield the same results. Often the relationship between "causes" and "effects" appears to be stronger in nonexperimental and quasi-experimental studies than in studies in which competing explanations are ruled out through randomization to different conditions.

For questions about prognosis or about etiology and harm (Table 2.1), both of which concern causal relationships, strong prospective (cohort) studies are usually the best design (although there are some situations in which etiology questions can involve randomization). Path analytic studies with longitudinal data and a strong theoretical basis can also be powerful. Retrospective case-control studies are relatively weak, by contrast. Systematic reviews of multiple prospective studies, together with support from theories or biophysiologic research, represent the strongest evidence for these types of question.

CRITIQUING GUIDELINES FOR STUDY DESIGN

The research design used in a quantitative study strongly influences the quality of its evidence and so should be carefully scrutinized. Researchers' design

BOX 9.1 Guidelines for Critiquing Research Designs in Quantitative Studies

1. What type of question (therapy, prognosis, etc.) is being addressed? Does the research question concern a possible causal relationship between the independent and dependent variables?
2. What would be the strongest design for the research question? How does this compare with the design actually used?
3. Is there an intervention or treatment? Was the intervention adequately described? Was the control or comparison condition adequately described? Was an experimental or quasi-experimental design used?
4. If the study was an RCT, what specific experimental design was used? Were randomization procedures adequately explained? Does the report provide evidence that randomization was successful—that is, resulted in groups that were comparable prior to the intervention? If cluster randomization was used, was there an adequate number of units?
5. If the design is quasi-experimental, what specific quasi-experimental design was used? Is there justification for deciding not to randomize participants to treatment conditions? Does the report provide evidence that any groups being compared were equivalent prior to the intervention?
6. If the design was nonexperimental, was the study inherently nonexperimental? If not, is there justification for not manipulating the independent variable? What specific nonexperimental design was used? If a retrospective design was used, is there justification for not using a prospective design? What evidence does the report provide that any groups being compared were similar with regard to important confounding characteristics?
7. What types of comparisons are specified in the design (e.g., before–after, between groups)? Do these comparisons adequately illuminate the relationship between the independent and dependent variables? If there are no comparisons, or faulty comparisons, how does this affect the study's integrity and the interpretability of the results?
8. Was the study longitudinal? Was the timing of the collection of data appropriate? Was the number of data collection points reasonable?
9. Was blinding/masking used? If yes, who was blinded—and was this adequate? If not, is there an adequate rationale for failure to mask? Is the intervention a type that could raise expectations that in and of themselves could alter the outcomes?

decisions have more of an impact on study quality than perhaps any other methodologic decision when the research question is about causal relationships.

Actual designs and some controlling techniques (randomization, blinding, allocation concealment) were described in this chapter, and the next chapter explains in greater detail specific strategies for enhancing research control. The guidelines in Box 9.1 are the first of two sets of questions to help you in critiquing quantitative research designs.

RESEARCH EXAMPLES

In this section, we present descriptions of an experimental, quasi-experimental, and nonexperimental study.

Research Example of an Experimental Study

Study: The Well Woman Program: A community-based randomized trial to prevent sexually transmitted infections in low-income African American women" (Marion et al., 2009).

Statement of Purpose: The purpose of the study was to determine the effectiveness of an intensive, culturally specific intervention designed to reduce sexually transmitted infections (STIs) among low-income African American women living in high-risk communities.

Treatment Groups: Nurse practitioners and trained peer educators delivered the Well Woman Program (WWP) in two phases. In the 2-month intensive phase, participants in the experimental group had a physical exam, received individual counseling, and attended group sessions led by peer educators. In the maintenance phase (months 3 through 12), they had ongoing tailored counseling and education. Participants in the "minimal intervention" control group received a 10-minute presentation on STIs, STI testing, and care as usual with community providers.

Method: A sample of 342 women from Chicago with a prior history of STIs was randomly assigned to the experimental or control group, using sealed envelopes with randomly generated numbers. Women were randomized in blocks of 10 to ensure comparable numbers in the two groups. Although study participants and those administering the intervention could not be blinded to the women's group status, data collectors were blinded. Data were collected from all women prior to random assignment and then at three follow-up points over the course of 15-months. The primary outcome was biologically confirmed sexually transmitted infection, using nucleic acid amplification tests on vaginal swabs. Participants also completed questionnaires with questions relating to STI risk behavior and other psychological variables.

Key Findings: Randomization appeared to be successful: the two groups were similar in terms of background characteristics that could affect STIs (e.g., age, number of lifetime partners), and in terms of baseline rate of having a positive test for an STI. At month 15, the estimated probability of WWP participants having an STI was 20% less than control group participants, leading the investigators to conclude that "better STI outcomes were due to the intensive individualized intervention" (p. 274).

Research Example of a Quasi-Experimental Study

Study: The impact of a multimedia informational intervention on healthcare service use among women and men newly diagnosed with cancer (Loiselle & Dubois, 2009).

Statement of Purpose: The purpose of the study was to test the effect of a comprehensive cancer informational intervention using information technology on patient satisfaction and the use of healthcare services by men and women newly diagnosed with cancer.

Treatment Groups: The intervention group received a 1-hour training session on the use of information technology, a CD-ROM with information on cancer, and a list of reputable cancer-related web sites. A research assistant was available by telephone or email to answer questions. Intervention materials (including laptop computers for those without a home computer) were available for an 8-week period. The control group received usual care.

Method: Patients from four cancer clinics within large teaching hospitals in Montreal were involved in this study. Eligible patients in three clinics were recruited into the intervention group, while those in the fourth clinic were recruited as the controls. To be eligible, patients had to be newly diagnosed with either breast or prostate cancer and had to plan cancer treatment in one of the study sites. Altogether, 250 patients agreed

to participate, 148 in the intervention group and 102 in the comparison group. Data relating to healthcare service use, patient satisfaction, perceptions of information support, and other variables were collected prior to the intervention, 9 weeks later, and then again 3 months later.

Key Findings: The intervention and comparison group members were similar demographically in some respects (e.g., marital status), but several preintervention group differences were found. For example, patients in the intervention group were younger and better educated than those in the comparison group. To address this selection bias problem, these characteristics were controlled statistically, an approach discussed in the next chapter. Patients in the two groups did not differ in their reliance on healthcare services following the intervention. However, patients in the experimental group were significantly more satisfied than those in the comparison group with the cancer information they received.

Research Example of a Correlational Study

Study: Placental position and late stillbirth: A case-control study (Warland, et al., 2009)

Statement of Purpose: The purpose of the study was to examine whether placental position in pregnancy contributes to the risk of having a stillbirth. Earlier research had suggested that some implantation sites may not provide adequate supply of nutrients and oxygen to the fetus.

Method: Pregnant women from two Australian obstetric hospitals were included in the sample. The cases were women with a discharge diagnosis of stillbirth who were at 27 or more weeks gestation. The control group comprised women who gave birth to a live baby at the same hospital during the same period. Controls were matched to cases on maternal age, infant gender, and gestational age. The researchers attempted to match two controls for every case, and were successful for all but five cases. Another nine cases could not be matched to any live-birth mother, and these were removed from the sample. The final sample consisted of 124 cases and 243 controls. The researchers retrospectively reviewed clinical records for all women and recorded the placental position that had been noted during a routine second trimester ultrasound.

Key Finding: Women who had a posterior located placenta were significantly more likely to suffer a stillbirth than women who had a placenta in any other position.

SUMMARY POINTS

- Many quantitative nursing studies aim to elucidate *cause-and-effect relationships*. The challenge of research design is to facilitate inferences about causality.
- Various criteria are used to establish causality. One criterion is that an observed relationship between a presumed cause (independent variable) and an effect (dependent variable) cannot be explained as being caused by other (confounding) variables.
- In an idealized model, a **counterfactual** is what would have happened to the same people simultaneously exposed *and* not exposed to the causal factor. The *effect* represents the difference between the two. The goal of research design is to find a good approximation to the idealized counterfactual.
- **Experiments** (or **randomized controlled trials [RCTs])** involve **manipulation** (the researcher manipulates the independent variable by introducing a **treatment** or **intervention**); control (including use of a **control group** that is not given the intervention and represents the comparative counterfactual); and **randomization** or **random assignment** (with people allocated to experimental and control groups at random to form groups that are comparable at the outset).
- Everyone in the experimental group usually gets the same intervention as delineated in formal protocols, but some studies involve **patient-centered interventions (PCIs)** that are tailored to meet individual needs or characteristics.
- Researchers can expose the control group to various conditions, including no treatment, an alternative treatment, a **placebo** or pseudointervention, standard treatment ("usual care"), different doses of the treatment, or a **wait-list** (*delayed treatment*) group.
- Random assignment is done by methods that give every participant an equal chance of being in any group, such as by flipping a coin or using

a **table of random numbers**. Randomization is the most reliable method for equating groups on all characteristics that could affect study outcomes. Randomization should involve **allocation concealment** that prevents foreknowledge of upcoming assignments.

- Randomization sometimes involves **stratification** in which participations are first divided into groups (e.g., men and women) before being randomized. In **permuted block randomization**, randomization is done for blocks of people—for example, 6 or 8 at a time in randomly selected block sizes—to ensure a balanced allocation to groups within cohorts of participants.
- **Blinding** (or **masking**) is sometimes used to avoid biases stemming from participants' or research agents' awareness of group status or study hypotheses. **Single-blind studies** involve masking of one group (e.g., participants) and **double-blind studies** involve masking of two groups (e.g., participants, investigators).
- The standard process is to randomize *individuals* to conditions after informed consent and the collection of **baseline data**, but there are variations. **Cluster randomization** involves randomizing larger units (e.g., hospitals) to treatment conditions. **Partially randomized patient preference (PRPP) designs** involve randomizing only patients without a treatment preference. **Randomized consent** (or **Zelen**) **designs** randomize prior to informed consent.
- A **posttest-only** (or **after-only**) **design** involves collecting data only after an intervention. In a **pretest–posttest** (or **before–after**) **design**, data are collected both before and after the intervention, permitting an analysis of change.
- **Factorial designs**, in which two or more independent variables are manipulated simultaneously, allow researchers to test both **main effects** (effects from manipulated independent variables) and **interaction effects** (effects from combining treatments).
- In a **crossover design**, people are exposed to more than one experimental condition, administered in a randomized order, and thus serve as their own controls.
- Experimental designs are the "gold standard" because they come closer than any other design in meeting criteria for inferring causal relationships.
- **Quasi-experimental designs** (*controlled trials without randomization*) involve an intervention but lack randomization. Strong quasi-experimental designs include features in support of causal inferences.
- The **nonequivalent control group pretest–posttest design** involves using a nonrandomized **comparison group** and the collection of pretreatment data so that initial group equivalence can be assessed. Comparability of groups can be sometimes be enhanced through *matching* on individual characteristics or by **propensity matching** that involves matching on a **propensity score** for each participant.
- In a **time series design**, there is no comparison group; information on the dependent variable is collected over a period of time before and after the intervention. Time series designs are often used in **single-subject (*N*-of-1) experiments**.
- Other quasi-experimental designs include the **regression discontinuity design**, quasi-experimental **dose-response analyses**, and the quasi-experimental (nonrandomized) **arms** of a PRPP randomization design (i.e., groups with strong preferences).
- In evaluating the results of quasi-experiments, it is important to ask whether it is plausible that factors other than the intervention caused or affected the outcomes (i.e., whether there are **rival hypotheses** for explaining the results).
- **Nonexperimental** (or **observational**) **research** includes **descriptive research**—studies that summarize the status of phenomena—and **correlational** studies that examine relationships among variables but involve no manipulation of the independent variable (often because it *cannot* be manipulated).
- Designs for correlational studies include **retrospective (case-control) designs** (which begin with the outcome and look back in time for antecedent causes of "caseness" by comparing **cases** that have a disease or condition with

controls who do not); **prospective (cohort) designs** (studies that begin with a presumed cause and look forward in time for its effect); **natural experiments** (in which a group is affected by a seemingly random event, such as a disaster); and **path analytic studies** (which test causal models developed on the basis of theory).

- **Descriptive correlational studies** describe how phenomena are interrelated without invoking a causal explanations. **Univariate descriptive studies** examine the frequency or average value of variables.
- Descriptive studies include **prevalence studies** that document the prevalence rate of a condition at one point in time and **incidence studies** that document the frequency of *new* cases, over a given time period. When the incidence rates for two groups are determined, it is possible to compute the **relative risk** of "caseness" for the two.
- The primary weakness of correlational studies for cause-probing questions is that they can harbor biases due to **self-selection** into groups being compared.

STUDY ACTIVITIES

Chapter 9 of the *Resource Manual for Nursing Research: Generating and Assessing Evidence for Nursing Practice, 9th ed.*, offers study suggestions for reinforcing concepts presented in this chapter. In addition, the following questions can be addressed in classroom or online discussions:

1. Assume that you have 10 people—Z, Y, X, W, V, U, T, S, R, and Q—who are going to participate in an RCT you are conducting. Using a table of random numbers, assign five individuals to group 1 and five to group 2.
2. Insofar as possible, use the questions in Box 9.1 to critique the three research examples described at the end of the chapter.
3. Discuss how you would design a prospective study to address the question posed in the Warland and colleagues (2009) case-control study summarized at the end of the chapter.

STUDIES CITED IN CHAPTER 9

Chen, S. Y., & Tzeng, Y. (2009). Path analysis for adherence to pelvic floor exercise among women with urinary incontinence. *Journal of Nursing Research, 17*, 83–92.

Colón-Emeric, C., Ammarell, N, Bailey, D., Corazzini, K., Lekan-Rutledge, D., Piven, M., Utley-Smith, Q., & Anderson, R. A. (2006). Patterns of medical and nursing staff communication in nursing homes. *Qualitative Health Research, 16*, 173–188.

Coward, D. (2002). Partial randomization design in a support group intervention study. *Western Journal of Nursing Research, 24*, 406–421.

Elliott, A., & Horgas, A. (2009). Effects of an analgesic trial in reducing pain behaviors in community-dwelling older adults with dementia. *Nursing Research, 58*, 140–145.

Hatfield, L. A. (2008). Sucrose decreases infant biobehavioral pain response to immunizations: A randomized controlled trial. *Journal of Nursing Scholarship, 40*, 219–225.

Heidrich, S., Brown, R., Egan, J., Perez, O., Phelan, C., Yeom, H., & Ward, S. (2009). An individualized representational intervention to improve symptom management (IRIS) in older breast cancer survivors. *Oncology Nursing Forum, 36*, E133–E143.

Hicks-Moore, S. (2005). Relaxing music at mealtime in nursing homes: Effects on agitated patients with dementia. *Journal of Gerontological Nursing, 31*, 26–32.

Huizing, A., Hamers, J., Gulpers, M., & Berger, M. (2009). Preventing the use of physical restraints on residents newly admitted to psycho-geriatric nursing home wards: A cluster-randomized trial. *International Journal of Nursing Studies, 46*, 459–469.

Jacob, E., Sockrider, M., Dinu, M., Acosta, M., & Mueller, B. (2010). Respiratory symptoms and acute painful episodes in sickle cell disease. *Journal of Pediatric Oncology Nursing, 27*, 33–39.

Johansson, Y., Bachrach-Lindstrom, M., Carstensen, J., & Ek, A. (2009). Malnutrition in a home-living older population: Prevalence, incidence and risk factors. *Journal of Clinical Nursing, 18*, 1354–1364.

Kratz, A. (2008). Use of the acute confusion protocol: A research utilization project. *Journal of Nursing Care Quality, 23*, 331–337.

Lai, H., & Good, M. (2005). Music improves sleep quality in older adults. *Journal of Advanced Nursing, 49*, 234–244.

Lai, H., Chen, C., Peng, T., Chang, F., Hsieh, M., Huang, H., & Chang, S. (2006). Randomized controlled trial of music during kangaroo care on maternal state anxiety and preterm infants' responses. *International Journal of Nursing Studies, 43*, 139–146.

Lauver, D., Settersten, L., Kane, J., & Henriques, J. (2003). Tailored messages, external barriers, and women's utilization

of professional breast cancer screening over time. *Cancer, 97*, 2724–2735.

Liehr, P., Mehl, M., Summers, L., & Pennebaker, J. (2004). Connecting with others in the midst of stressful upheaval on September 11, 2001. *Applied Nursing Research, 17*, 2–9.

Loiselle, C., & Dubois, S. (2009). The impact of a multimedia informational intervention on healthcare service use among women and men newly diagnosed with cancer. *Cancer Nursing, 32*, 37–44.

Marion, L. N., Finnegan, L., Campbell, R., & Szalacha, L. (2009). The Well Woman Program: A community-based randomized trial to prevent sexually transmitted infections in low-income African American women. *Research in Nursing & Health, 32*, 274–285.

Martinez, C., Carmelli, E., Barak, S., & Stopka, C. (2009). Changes in pain-free walking based on time in accommodating pain-free exercise therapy for peripheral artery disease. *Journal of Vascular Nursing, 27*, 2–7.

Munro, C., Grap, M., Jones, D., McClish, D., & Sessler, C. (2009). Chlorhexidine, toothbrushing, and preventing ventilator-assisted pneumonia in critically ill adults. *American Journal of Critical Care, 18*, 428–437.

Musil, C., Warner, C., Zauszniewski, J., Wykle, M., & Standing, T. (2009). Grandmother caregiving, family stress and strain, and depressive symptoms. *Western Journal of Nursing Research, 31*, 389–408.

Nikolajsen, L., Lyndgaard, K., Schriver, N., & Mollet, J. (2009). Does audiovisual stimulation with music and nature sights reduce pain and discomfort during placement of a femoral nerve block? *Journal of Perianesthesia Nursing, 24*, 14–18.

Pinar, K., Moore, K., Smits, E., Murphy, K., & Schopflocher, D. (2009). Leg bag comparison: Reported skin health, comfort, and satisfaction. *Journal of Wound, Ostomy, & Continence Nursing, 36*, 319–326.

Pölkki, T., Pietilä, A., Vehviläinen-Julkunen, K., Laukkala, H., & Kiviluoma, K. (2008). Imagery-induced relaxation in children's postoperative pain relief. *Journal of Pediatric Nursing, 23*, 217–224.

Seers, K., Crichton, N., Tutton, L., Smith, L., & Saunders, T. (2008). Effectiveness of relaxation for postoperative pain and anxiety: randomized controlled trial. *Journal of Advanced Nursing, 62*, 681–688.

Steiner, A., Walsh, B., Pickering, R., Wiles, R., Ward, J., Brooking, J., & Southampton NLU Evaluation Team. (2001). Therapeutic nursing or unblocking beds? A randomized controlled trial of a post-acute intermediate care unit. *British Medical Journal, 322*(7284), 453–460.

Swadener-Culpepper, L., Skaggs, R., & Vangilder, C. (2008). The impact of continuous lateral rotation therapy in overall clinical and financial outcomes of critically ill patients. *Critical Care Nursing Quarterly, 31*, 270–279.

Swenson, K., Nissen, M., Leach, J., & Post-White, J. (2009). Case-control study to evaluate predictors of lymphedema after breast cancer surgery. *Oncology Nursing Forum, 36*, 185–193.

Warland, J., McCutcheon, H., & Baghurst, P. (2009). Placental position and late stillbirth: A case control study. *Journal of Clinical Nursing, 18*, 1602–1606.

Wentworth, L., Briese, L., Timimi, F., Bartel, D., Cutshall, S., Tilbury, R., Lennon, R., & Bauer, B. A. (2009). Massage therapy reduces tension, anxiety, and pain in patients awaiting invasive cardiovascular procedures. *Progress in Cardiovascular Nursing, 24*, 155–161.

Wiklund, I., Edman, G., Larsson, C., & Andolf, E. (2009). First-time mothers and changes in personality in relation to mode of delivery. *Journal of Advanced Nursing, 65*, 1636–1644.

Yuan, S., Chou, M., Hwu, L., Chang, Y., Hsu, W., & Kuo, H. (2009). An intervention program to promote health-related physical fitness in nurses. *Journal of Clinical Nursing, 18*, 1404–1411.

Methodologic and nonresearch references cited in this chapter can be found in a separate section at the end of the book.

10 Rigor and Validity in Quantitative Research

VALIDITY AND INFERENCE

This chapter describes strategies for enhancing the rigor of quantitative studies, including ways to minimize biases and control confounding variables. Most of these strategies help to strengthen the inferences that can be made about cause-and-effect relationships.

Validity and Validity Threats

In designing a study, a constructive approach is to anticipate the possible factors that could undermine the **validity** of inferences. Shadish and colleagues (2002) define validity in the context of research design as "the approximate truth of an inference" (p. 34). For example, inferences that an *effect* results from a hypothesized *cause* are valid to the extent that researchers can marshal supporting evidence. Validity is always a matter of degree, not an absolute.

Validity is a property of an inference, not of a research design, but design elements profoundly affect the inferences that can be made. **Threats to validity** are reasons that an inference could be wrong. When researchers introduce design features to minimize potential threats, the validity of the inference is strengthened, and thus evidence is more persuasive. We identify important validity threats to encourage you to think about ways to address them during the design phase of a study and to evaluate them in interpreting study results.

Types of Validity

Shadish and colleagues (2002) proposed a validity taxonomy that identified four aspects of a good research design, and catalogued dozens of threats to validity. This chapter describes the taxonomy and briefly summarizes major threats, but we urge researchers to consult this seminal work for further guidance on strengthening study validity.

The first type of validity, **statistical conclusion validity**, concerns the validity of inferences that there truly is an empirical relationship, or correlation, between the presumed cause and the effect. The researcher's job is to provide the strongest possible evidence that the relationship is *real* and that the intervention (if any) was given a fair test.

Internal validity concerns the validity of inferences that, given that an empirical relationship exists, it is the independent variable, rather than something else, that caused the outcome. The researcher's job is to develop strategies to rule out the plausibility that something other than the independent variable accounts for the observed relationship.

Construct validity involves the validity of inferences "from the observed persons, settings, and cause-and-effect operations included in the study to the constructs that these instances might represent" (p. 38). One aspect of construct validity concerns the degree to which an intervention is a good representation of the underlying construct that was theorized as having the potential to cause beneficial outcomes. Another concerns whether the measures of the dependent variable are good operationalizations of the constructs for which they are intended.

External validity concerns whether inferences about observed relationships will hold over variations in persons, setting, time, or measures of the outcomes. External validity, then, is about the generalizability of causal inferences, and this is a critical concern for research that aims to yield evidence for evidence-based nursing practice.

These four types of validity and their associated threats are discussed in this chapter. Many validity threats concern inadequate control over confounding variables, so we briefly review methods of controlling variation associated with characteristics of study participants.

Controlling Intrinsic Source of Confounding Variability

This section describes six ways of controlling confounding participant characteristics to rule out rival explanations for cause-and-effect relationships.

Randomization

Randomization is the most effective method of controlling individual characteristics. The primary function of randomization is to secure comparable groups—that is, to equalize groups with respect to confounding variables. A distinct advantage of random assignment, compared with other control methods, is that it controls *all* possible sources of extraneous variation, *without any conscious decision about which variables need to be controlled.*

Crossover

Randomization within a crossover design is an especially powerful method of ensuring equivalence between groups being compared—participants serve as their own controls. Moreover, fewer participants usually are needed in such a design. Fifty people exposed to two treatments in random order yield 100 pieces of data (50 × 2); 50 people randomly assigned to two different groups yield only 50 pieces of data (25 × 2). Crossover designs are not appropriate for all studies, however, because of the possible carry-over effects: People exposed to two different conditions may be influenced in the second condition by their experience in the first.

Homogeneity

When randomization and crossover are not feasible, alternative methods of controlling confounding characteristics are needed. One method is to use only people who are homogeneous with respect to confounding variables—that is, confounding traits are not allowed to vary. Suppose we were testing the effectiveness of a physical fitness program on the cardiovascular functioning of elders. Our quasi-experimental design involves elders from two different nursing homes, with elders in one of them receiving the physical fitness program. If gender were an important confounding variable (and if the two nursing homes had different proportions of men and women), we could control gender by using only men (or only women) as participants.

Using a homogeneous sample is easy as a control mechanism, but the price is that research findings can be generalized only to the type of people who participated in the study. If the physical fitness program were found to have beneficial effects on the cardiovascular status of a sample of women 65 to 75 years of age, its usefulness for improving the cardiovascular status of men in their 80s would require a separate study. Indeed, one noteworthy criticism of this approach is that researchers sometimes exclude people who are extremely ill, which means that the findings cannot be generalized to those who perhaps are most in need of interventions.

Example of control through homogeneity: Ngai and colleagues (2010) studied factors that predicted maternal role competence and satisfaction among mothers in Hong Kong. Several variables were controlled through homogeneity, including ethnicity (all were Chinese), parity (all primiparous), and marital status (all were married).

TIP: The principle of homogeneity is often used to control (hold constant) external factors as well as participant characteristics. For example, it may be important to collect outcome data at the same time of the day for all participants if time could affect the outcome (e.g., fatigue). As another example, it may be desirable to maintain **constancy of conditions** in terms of locale of data collection—for example, interviewing all respondents in their own homes, rather than some in their places of work. In each setting, participants assume different roles (e.g., spouse and parent versus employee), and responses may be influenced to some degree by those roles.

Stratification/Blocking

Another approach to controlling confounding variables is to include them in the research design through stratification, as discussed in Chapter 9. To pursue our example of the physical fitness program with gender as the confounding variable, we could build it into the study in a randomized block design in which elderly men and women would be randomly assigned separately to treatment groups. This approach can enhance the likelihood of detecting differences between our experimental and control groups because we can eliminate the effect of the blocking variable (gender) on the dependent variable. In addition, if the blocking variable is of interest substantively, this approach gives researchers the opportunity to study differences in groups created by the stratifying variable (e.g., men versus women). Stratification is appropriate in experiments, and is used in quasi-experimental and correlational studies as well.

Matching

Matching (also called **pair matching**) involves using information about people's characteristics to create comparable groups. If matching were used in our physical fitness example, and age and gender were the confounding variables, we would match a person in the program group with one in the comparison group with respect to age and gender. As noted in the previous chapter, there are reasons why matching is problematic. First, to use matching, researchers must know the relevant confounding variables in advance. Second, it is often difficult to match on more than two or three variables, unless propensity score matching is used—but this method requires technical sophistication. Yet there are usually many confounding variables that could affect outcomes of interest. For these reasons, matching as the primary control technique should be used only when other, more powerful procedures are not feasible, as might be the case in some nonexperimental studies (e.g., case-control designs).

Sometimes, as an alternative to pair matching, researchers use a *balanced design* with regard to key confounders. In such situations, researchers attempt only to ensure that the groups being compared have similar proportional representation on confounding variables, rather than matching on a one-to-one basis. For example, if gender and age were the two variables of concern, we would strive to ensure that the same percentage of men and women were in the two groups and that the average age was comparable. Such an approach is less cumbersome than pair matching, but has similar limitations. Nevertheless, both pair matching and balancing are preferable to failing to control participant characteristics at all.

Example of control through matching: Luttik and colleagues (2009) studied quality of life in partners of people with congestive heart failure, in comparison to those living with a healthy partner. The two groups of partners were matched in terms of gender and age.

Statistical Control

Another method of controlling confounding variables is through statistical analysis rather than research design. A detailed description of powerful **statistical control** mechanisms will be postponed until Chapter 18, but we will explain underlying principles with a simple illustration of a procedure called **analysis of covariance (ANCOVA)**.

In our physical fitness example, suppose we used a nonequivalent control group design with residents from two nursing homes, and resting heart rate was an outcome. We would expect individual differences in heart rate within the sample—that is, it would vary from one person to the next. The research question is, Can some of the individual differences in heart rate be attributed to a person's participation in physical fitness? We know that differences in heart rate are also related to other characteristics, such as age. In Figure 10.1, the large circles represent the

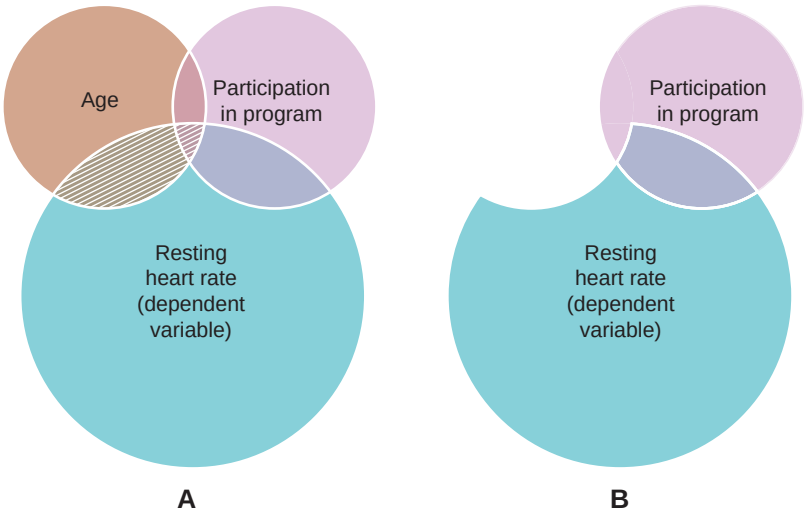


FIGURE 10.1 Schematic diagram illustrating the principle of analysis of covariance.

total extent of individual differences for resting heart rate. A certain amount of variability can be explained by a person's age, which is the small circle on the left in Figure 10.1A. Another part of the variability can perhaps be explained by participation or nonparticipation in the program, represented as the small circle on the right. The two small circles (age and program participation) overlap, indicating a relationship between the two. In other words, people in the physical fitness group are, on average, either older or younger than those in the comparison group, and so age should be controlled. Otherwise, it will be impossible to determine whether postintervention differences in resting heart rate are attributable to differences in age or program participation.

Analysis of covariance controls by statistically removing the effect of confounding variables on the outcome. In the illustration, the portion of heart rate variability attributable to age (the hatched area of the large circle in A) is removed through ANCOVA. Figure 10.1B shows that the final analysis assesses the effect of program participation on heart rate *after removing the effect of age*. By controlling heart rate variability resulting from age, we get a more accurate estimate of the effect of the program on heart rate. Note that even after removing variability due to age, there is still individual variation not associated with the program treatment—the bottom half of the large circle in B. This means that the study can probably be further enhanced by controlling additional confounders that might account for heart rate differences in the two nursing homes, such as gender, smoking history, and so on. Analysis of covariance and other sophisticated procedures can control multiple confounding variables.

Example of statistical control: Lee and colleagues (2009) tested the effectiveness of a 26-week Tai Chi intervention on health-related quality of life (QOL) in residents from six nursing homes, two of which got the intervention and the other four of which did not. Changes in QOL for residents receiving and not receiving the intervention were compared, while controlling statistically for resident satisfaction.

TIP: Confounding participant characteristics that need to be controlled vary from one study to another, but we can offer some guidance. The best variable is the dependent variable itself, measured before the independent variable occurs. In our physical fitness example, controlling preprogram measures of cardiovascular functioning through ANCOVA would be especially powerful because this would remove the effect of individual variation stemming from many other extraneous factors. Major demographic variables (e.g., age, race/ethnicity, education) and health status indicators are usually good candidates to measure and control. Confounding variables that need to be controlled—variables that correlate with the outcomes—should be identified through a literature review.

Evaluation of Control Methods

Table 10.1 summarizes benefits and drawbacks of the six control mechanisms. Randomization is the most effective method of managing confounding variables—that is, of approximating the ideal but unattainable counterfactual discussed in Chapter 9—because it tends to cancel out individual differences on all possible confounders. Crossover designs are a useful supplement to randomization, but are not always appropriate. The remaining alternatives have a common disadvantage: Researchers must know in advance the relevant confounding variables. To

TABLE 10.1 Methods of Control over Participant Characteristics

METHOD	BENEFITS	LIMITATIONS
Randomization	• Controls all preintervention confounding variables • Does not require advance knowledge of which variables to control	• Ethical and practical constraints on variables that can be manipulated • Possible artificiality of conditions
Crossover	• If done with randomization, strongest possible approach	• Cannot be used if there are possible carry-over effect from one condition to the next • History threat may be relevant if external factors change over time
Homogeneity	• Easy to achieve in all types of research • Could enhance interpretability of relationships	• Limits generalizability • Requires knowledge of which variables to control • Range restriction could lower statistical conclusion validity
Stratification	• Enhances the ability to detect and interpret relationships • Offers opportunity to examine blocking variable as an independent variable	• Usually restricted to a few stratifying variables • Requires knowledge of which variables to control
Matching	• Enhances ability to detect and interpret relationships • May be easy if there is a large "pool" of potential available controls	• Usually restricted to a few matching variables (except with propensity matching) • Requires knowledge of which variables to match • May be difficult to find comparison group matches, especially if there are more than two matching variables
Statistical control	• Enhances ability to detect and interpret relationships • Relatively economical means of controlling several confounding variables	• Requires knowledge of which variables to control, as well as measurement of those variables • Requires some statistical sophistication

select homogeneous samples, stratify, match, or perform ANCOVA, researchers must know which variables need to be measured and controlled. Yet, when randomization is impossible, the use of any of these strategies is better than no control strategy at all.

STATISTICAL CONCLUSION VALIDITY

As noted in Chapter 9, one criterion for establishing causality is demonstrating that there is a relationship between the independent and dependent variable. Statistical methods are used to support inferences about whether relationships exist. Design decisions can influence whether statistical tests will detect true relationships, so researchers need to make decisions that protect against reaching false statistical conclusions. Even for research that is not cause probing, researchers need to attend to statistical conclusion validity: The issue is whether relationships that exist in reality can be reliably detected in a study. Shadish and colleagues (2002) discussed nine threats to statistical conclusion validity. We focus here on three especially important threats.

Low Statistical Power

Statistical power refers to the ability to detect true relationships among variables. Adequate statistical power can be achieved in various ways, the most straightforward of which is to use a sufficiently large sample. When small samples are used, statistical power tends to be low, and the analyses may fail to show that the independent and dependent variables are related—*even when they are*. Power and sample size are discussed in Chapters 12 and 17.

Another aspect of a powerful design concerns how the independent variable is defined. Both statistically and substantively, results are clearer when differences between groups being compared are large. Researchers should aim to maximize group differences on the dependent variables by maximizing differences on the independent variable. Conn and colleagues (2001) offer good suggestions for enhancing the power and effectiveness of nursing interventions. Strengthening group differences is usually easier in experimental than in nonexperimental research. In experiments, investigators can devise treatment conditions that are as distinct as money, ethics, and practicality permit. Even in nonexperimental research, however, there may be opportunities to operationalize independent variables in such a way that power to detect differences is enhanced.

Another aspect of statistical power concerns maximizing **precision**, which is achieved through accurate measuring tools, controls over confounding variables, and powerful statistical methods. Precision can best be explained through an example. Suppose we were studying the effect of admission into a nursing home on depression by comparing elders who were or were not admitted. Depression varies from one elderly person to another for various reasons. We want to isolate—as precisely as possible—the portion of variation in depression attributable to nursing home admission. Mechanisms of research control that reduce variability attributable to confounding factors can be built into the research design, thereby enhancing precision. The following ratio expresses what we wish to assess in this example:

$$\frac{\text{Variability in depression due to nursing home admission}}{\text{Variability in depression due to other factors (e.g., age, pain, medical condition)}}$$

This ratio, greatly simplified here, captures the essence of many statistical tests. We want to make variability in the numerator (the upper half) as large as possible relative to variability in the denominator (the lower half), to evaluate precisely the relationship between nursing home admission and depression. The smaller the variability in depression due to confounding variables (e.g., age, pain), the easier it will be to detect differences in depression between elders who

were or were not admitted to a nursing home. Designs that enable researchers to reduce variability caused by confounders can increase statistical conclusion validity. As a purely hypothetical illustration, we will attach some numeric values* to the ratio as follows:

$$\frac{\text{Variability due to nursing home admission}}{\text{Variability due to all confounding variables}} = \frac{10}{4}$$

If we can make the bottom number smaller, say by changing it from 4 to 2, we will have a more precise estimate of the effect of nursing home admission on depression, relative to other influences. Control mechanisms such as those described earlier help to reduce variability caused by extraneous variables and should be considered as design options in planning a study. We illustrate this by continuing our example, singling out age as a key confounding variable. Total variability in levels of depression can be conceptualized as having the following components:

$$\text{Total variability in depression} = \text{Variability due to nursing home admission} + \text{Variability due to age} + \text{Variability due to other confounding variables}$$

This equation can be taken to mean that part of the reason why some elders are depressed and others are not is that some were admitted to a nursing home and others were not; some were older and some were younger; other factors, such as level of pain and medical condition, also had an effect on depression.

One way to increase precision in this study would be to control age, thereby removing the variability in depression that results from age differences. We could do this, for example, by restricting age to elders younger than 80, thereby reducing the variability in depression due to age. As a result, the effect of nursing home admission on depression becomes greater, relative to the remaining variability. Thus, this design decision (homogeneity) enabled us to get a more precise estimate of the effect of nursing home admission on level of depression (although, of course, this limits generalizability). Research designs differ considerably in the sensitivity with which effects under study can be detected statistically. Lipsey (1990) has prepared an excellent guide to assist researchers in enhancing the sensitivity of research designs.

Restriction of Range

Although the control of extraneous variation through homogeneity is easy to use and can help to clarify the relationship between key research variables, it can be risky. Not only does this approach limit the generalizability of study findings, but it can also sometimes undermine statistical conclusion validity. When the use of homogeneity restricts the range of values on the outcome variable, relationships between the outcome and the independent variable will be *attenuated*, and may, therefore, lead to an erroneous inference that the variables are unrelated.

In the example just used, we suggested limiting the sample of nursing home residents to elders younger than 80 to reduce variability in the denominator. Our aim was to enhance the variability in depression scores attributable to nursing home admission, relative to depression variability due to other factors. What if, however, few elders under 80 were depressed? With limited variability, relationships cannot be detected—the values in both the numerator and denominator are deflated. For example, if *everyone* had a depression score of 50, depression scores would be totally unrelated to age, pain levels, nursing home admission, and so on. Thus, in designing a study, it is important to consider whether there will be sufficient variability to support the statistical analyses envisioned. The issue of *floor effects* and *ceiling effects*, which involve range restrictions at the lower and upper end of a measure, respectively, are discussed later in this book.

*You should not be concerned with how these numbers can be obtained. Analytic procedures are explained in Chapter 17.

TIP: In designing a study, try to anticipate nonsignificant findings, and consider design adjustments that might affect the results. For example, suppose our study hypothesis is that environmental factors such as light and noise affect acute confusion in the hospitalized elderly. With a preliminary design in mind, imagine findings that *fail* to support the hypothesis. Then ask yourself what could be done to decrease the likelihood of getting such negative results, under the assumption that such results do not reflect the truth. Could power be increased by making differences in environmental conditions sharper? Could precision be increased by controlling additional confounding variables? Could bias be eliminated by better training of research staff?

Unreliable Implementation of a Treatment

The strength of an intervention (and hence statistical conclusion validity) can be undermined if an intervention is not as powerful in reality as it is "on paper." **Intervention fidelity (**or **treatment fidelity)** concerns the extent to which the implementation of an intervention is faithful to its plan. There is growing interest in intervention fidelity in the nursing literature and considerable advice on how to achieve it (e.g., Spillane et al., 2007; Stein et al., 2007; Whitmer et al., 2005).

Interventions can be weakened by various factors, which researchers can often influence. One issue concerns the extent to which the intervention is similar from one person to the next. Usually, researchers strive for constancy of conditions in implementing a treatment because lack of standardization adds extraneous variation and can diminish the intervention's full force. Even in tailored, patient-centered interventions there are usually protocols, though different protocols are used with different people. Using the notions just described, when standard protocols are not followed, variability due to the intervention (i.e., in the numerator) can be suppressed, and variability due to other factors (i.e., in the denominator) can be inflated, possibly leading to the erroneous conclusion that the intervention was ineffective. This suggests the need for a certain degree of standardization, the development of procedures manuals, thorough training of personnel, and vigilant monitoring (e.g., through observations of the delivery of the intervention) to ensure that the intervention is being implemented as planned—and that control group members have not gained access to the intervention.

Determining that the intervention was delivered as intended may need to be supplemented with efforts to ensure that the intervention was *received* as intended. This may involve a **manipulation check** to assess whether the treatment was in place, was understood, or was perceived in an intended manner. For example, if we were testing the effect of soothing versus jarring music on anxiety, we might want to determine whether participants themselves perceived the music as soothing and jarring. Another aspect of treatment fidelity for interventions designed to promote behavioral changes concerns the concept of *enactment* (Bellg et al., 2004). Enactment refers to participants' performance of the treatment-related skills, behaviors, and cognitive strategies in relevant real-life settings.

Example of attention to treatment fidelity: Radziewicz and colleagues (2009) described their efforts to establish treatment fidelity in a telephone intervention to provide support to aging patients with cancer and their family caregivers. Their treatment fidelity plan included monitoring adherence to standards of a protocol, carefully training staff using a standardized manual, monitoring the success of training, and monitoring consistency in delivering the intervention.

Another issue is that participants often fail to receive the desired intervention due to lack of **treatment adherence**. It is not unusual for those in the experimental group to elect not to participate fully in the treatment—for example, they may stop going to treatment sessions. To the extent possible, researchers should take steps to encourage participation among those in the treatment group. This might mean making the intervention as enjoyable as possible, offering incentives, and reducing burden in terms of the intervention and data collection (Polit & Gillespie, 2010). Nonparticipation in an intervention is rarely random, so researchers should document which people got what amount of treatment so that individual differences in "dose" can be taken into account in the analysis or interpretation of results.

TIP: Except for small-scale studies, every study should have a **procedures manual** that delineates the protocols and procedures for its implementation. The Toolkit section of the accompanying *Resource Manual* provides a model table of contents for such a procedures manual. The Toolkit also includes a model checklist to monitor delivery of an intervention through direct observation of intervention sessions.

INTERNAL VALIDITY

Internal validity refers to the extent to which it is possible to make an inference that the independent variable, rather than another factor, is truly causing variation in the dependent variable. We infer from an effect to a cause by eliminating (controlling) other potential causes. The control mechanisms reviewed earlier are strategies for improving internal validity. If researchers do not carefully manage extraneous variation, the conclusion that participants' performance on the outcome was caused by the independent variable is open to challenge.

Threats to Internal Validity

True experiments possess a high degree of internal validity because manipulation and random assignment allows researchers to rule out most alternative explanations for the results. Researchers who use quasi-experimental or correlational designs must contend with competing explanations of what caused the outcomes. Major competing explanations, or threats to internal validity, are examined in this section.

Temporal Ambiguity

As noted in Chapter 9, a criterion for inferring a causal relationship is that the cause must precede the effect. In RCTs, researchers themselves create the independent variable and then observe subsequent performance on an outcome variable, so establishing temporal sequencing is never a problem. In correlational studies, however, it may be unclear whether the independent variable preceded the dependent variable, or vice versa.

Selection

Selection (self-selection) encompasses biases resulting from pre-existing differences between groups. When individuals are not assigned to groups randomly, the groups being compared could be nonequivalent. Differences on outcomes could then reflect group differences rather than the effect of the independent variable. For example, if we found that women with an infertility problem were more likely to be depressed than women who were mothers, it would be impossible to conclude that the two groups differed in depression *because* of childbearing differences; women in the two groups might have been different in psychological well-being from the start. The problem of selection is reduced if researchers can collect data on participants' characteristics before the occurrence of the independent variable. In our example, the best design would be to collect data on women's depression before they attempted to become pregnant, and then design the study to control early levels of depression. Selection bias is one of the most problematic and frequently encountered threats to the internal validity of studies not using an experimental design.

History

The threat of **history** refers to the occurrence of external events that take place concurrently with the independent variable, and that can affect the outcomes. For example, suppose we were studying the effectiveness of a nurse-led outreach program to encourage pregnant women in rural areas to improve health practices (e.g., cessation of smoking, earlier prenatal care). The program might be evaluated by comparing the average birth weight of infants born in the 12 months before the outreach program with the average birth weight of those born in the 12 months after the program was introduced, using a time series design. However, suppose that 1 month after the new program was launched, a well-publicized docudrama about the inadequacies of prenatal care for poor women was aired on television. Infants' birth weight might now be affected by both the intervention and the messages in the docudrama, and it becomes impossible to disentangle the two effects.

In a true experiment, history usually is not a threat to a study's internal validity because we can often assume that external events are as likely to affect the experimental as the control group. When this is the case, group differences on the dependent variables represent effects over and above those created by outside factors. There are, however, exceptions. For example, when a crossover design is used, an event external to the study may occur during the first half (or second half) of the experiment, so treatments would be contaminated by the effect of that event. That is, some people would receive treatment A with the event and others would receive treatment A without it, and the same would be true for treatment B.

Selection biases sometimes interact with history to compound the threat to internal validity. For example, if the comparison group is different from the treatment group, then the characteristics of the members of the comparison group could lead them to have different intervening experiences, thereby introducing both history and selection biases into the design.

Maturation

In a research context, **maturation** refers to processes occurring within participants during the course of the study as a result of the passage of time rather than as a result of the independent variable. Examples of such processes include physical growth, emotional maturity, and fatigue. For instance, if we wanted to evaluate the effects of a sensorimotor program for developmentally delayed children, we would have to consider that progress occurs in these children even without special assistance. A one-group pretest—posttest design, for example, is highly susceptible to this threat.

Maturation is often a relevant consideration in nursing research. Remember that maturation here does not refer just to aging, but rather to any change that occurs as a function of time. Thus, maturation in the form of wound healing, postoperative recovery, and other bodily changes could be a rival explanation for the independent variable's effect on outcomes.

Mortality/Attrition

Mortality is the threat that arises from attrition in groups being compared. If different kinds of people remain in the study in one group versus another, then these differences, rather than the independent variable, could account for observed differences on the dependent variables at the end of the study. The most severely ill patients might drop out of an experimental condition because it is too demanding, or they might drop out of the comparison group because they see no advantage to remaining in the study. In a prospective cohort study, there may be differential attrition between groups being compared because of death, illness, or geographic relocation. Attrition bias essentially is a type of selection bias that occurs after the unfolding of the study: Groups initially equivalent can lose comparability because of attrition, and it could be that the differential composition, rather than the independent variable, is the "cause" of any group differences on the dependent variables. Attrition bias can also occur in single-group quasi-experiments if those dropping out of the study are a biased subset that make it look like a change in average values resulted from a treatment.

The risk of attrition is especially great when the length of time between points of data collection is long. A 12-month follow-up of participants, for example, tends to produce higher rates of attrition than a 1-month follow-up (Polit & Gillespie, 2009). In clinical studies, the problem of attrition may be especially acute because of patient death or disability.

If attrition is random (i.e., those dropping out of a study are comparable to those remaining in it), then there would not be bias. However, attrition is rarely random. In general, the higher the rate of attrition, the greater the likelihood of bias.

TIP: In longitudinal studies, attrition may occur because researchers cannot find participants, rather than because they refused to stay in the study. One effective strategy to help **tracing** people is to obtain **contact information** from participants at each point of data collection. Contact information should include the names, addresses, and telephone numbers of two or three people with whom the participant is close (e.g., parents, close friends)—people who would be likely to know how to contact participants if they moved. A sample contact information form that can be adapted for your use is provided in the Toolkit of the accompanying Resource Manual.

Testing and Instrumentation

Testing refers to the effects of taking a pretest on people's performance on a posttest. It has been found, particularly in studies dealing with attitudes, that the mere act of collecting data from people changes them. Suppose a sample of nursing students completed a questionnaire about attitudes toward assisted suicide. We then teach them about various arguments for and against assisted suicide, outcomes of court cases, and the like. At the end of instruction, we give them the same attitude measure and observe whether their attitudes have changed. The problem is that the first questionnaire might sensitize students, resulting in attitude changes regardless of whether instruction follows. If a comparison group is not used, it becomes impossible to segregate the effects of the instruction from the effects of the pretest. Sensitization, or testing, problems are more likely to occur when pretest data are gathered via self-reports (e.g., in a questionnaire), especially if people are exposed to controversial or novel material in the pretest.

Another related threat is **instrumentation**. This bias reflects changes in measuring instruments or methods of measurement between two points of data collection. For example, if we used one measure of stress at baseline and a revised measure at follow-up, any differences might reflect changes in the measuring tool rather than the effect of an independent variable. Instrumentation effects can occur even if the same measure is used. For example, if the measuring tool yields more accurate measures on a second administration (e.g., if data collectors are more experienced) or less accurate measures the second time (e.g., if participants become bored and answer haphazardly), then these differences could bias the results.

Internal Validity and Research Design

Quasi-experimental and correlational studies are especially susceptible to threats to internal validity. Table 10.2 lists specific designs that are *most* vulnerable to the threats just described—although it should not be assumed that threats are irrelevant in designs not listed. Each threat represents an alternative explanation that competes with the independent variable as a cause of the dependent variable. The aim of a strong research design is to rule out competing explanations. (Tables 9.5 and 9.6 in Chapter 9 also include information about internal validity threats for specific designs.)

TABLE 10.2 Research Designs and Threats to Internal Validity

THREAT	DESIGNS MOST SUSCEPTIBLE
Temporal Ambiguity	Case-control Other retrospective/cross-sectional
Selection	Nonequivalent control group (especially, posttest-only) Case-control "Natural" experiments with two groups Time series, if the population changes over time
History	One-group pretest–posttest Time series Prospective cohort Crossover
Maturation	One-group pretest–posttest
Mortality/ Attrition	Prospective cohort Longitudinal experiments and quasi-experiments One-group pretest–posttest
Testing	All pretest–posttest designs
Instrumentation	All pretest–posttest designs

An experimental design normally rules out most rival hypotheses, but even in RCTs, researchers must exercise caution. For example, if there is treatment infidelity or contamination between treatments, then history might be a rival explanation for any group differences (or lack of differences). Mortality can be a salient threat in true experiments. Because the experimenter does things

differently with the experimental and control groups, people in the groups may drop out of the study differentially. This is particularly apt to happen if the experimental treatment is painful, inconvenient, or time-consuming or if the control condition is boring or bothersome. When this happens, participants remaining in the study may differ from those who left in important ways, thereby nullifying the initial equivalence of the groups.

In short, researchers should consider how best to guard against and detect all possible threats to internal validity, no matter what design is used.

Internal Validity and Data Analysis

The best strategy for enhancing internal validity is to use a strong research design that includes control mechanisms and design features discussed in this chapter. Even when this is possible (and, certainly, when this is *not* possible), it is advisable to conduct analyses to assess the nature and extent of biases. When biases are detected, the information can be used to interpret substantive results. And, in some cases, biases can be statistically controlled.

Researchers need to be self-critics. They need to consider fully and objectively the types of biases that could have arisen—and then systematically search for evidence of their existence (while hoping, of course, that no evidence can be found). To the extent that biases can be ruled out or controlled, the quality of evidence the study yields will be strengthened.

Selection biases should always be examined. Typically, this involves comparing groups on pretest measures, when pretest data have been collected. For example, if we were studying depression in women who delivered a baby by cesarean delivery versus those who delivered vaginally, selection bias could be assessed by comparing depression in these two groups during or before the pregnancy. If there are significant predelivery differences, then any postdelivery differences would have to be interpreted with initial differences in mind (or with differences controlled). In designs with no pretest measure of the outcome, researchers should assess selection biases by comparing groups with respect to key background variables such as age, health status, and so on. Selection biases should be analyzed even in RCTs because there is no guarantee that randomization will yield perfectly equivalent groups.

Whenever the research design involves multiple points of data collection, researchers should analyze attrition biases. This is typically achieved through a comparison of those who did and did not complete the study with regard to baseline measures of the dependent variable or other characteristics measured at the first point of data collection.

Example of assessing attrition and selection bias: Resnick and colleagues (2008) used a cluster-randomized design to study the effectiveness of an intervention to enhance the self-efficacy of minority urban-dwelling elders. At the 15-week follow-up, only 62% of the initial participants provided outcome data. Dropouts did not differ from those who completed the study in terms of baseline characteristics (attrition bias), and those in the experimental and control group were also similar at baseline (selection bias).

When people withdraw from an intervention study, researchers are in a dilemma about whom to "count" as being "in" a condition. A procedure that is often used is a **per-protocol analysis**, which includes members in a treatment group only if they actually received the treatment. Such an analysis is problematic, however, because self-selection into a nonintervention condition could undo the initial comparability of groups. This type of analysis will almost always be biased toward finding positive treatment effects. The "gold standard" approach is to use an **intention-to-treat analysis**, which involves keeping participants who were randomized in the groups to which they were assigned (Polit & Gillespie, 2009, 2010). An intention-to-treat analysis may yield an underestimate of the effects of a treatment if many participants did not actually get the assigned treatment—but may be a better reflection of what would happen in the real world. Of course, one difficulty with an intention-to-treat analysis is that it is often difficult to obtain outcome data for people who have dropped out of a treatment, but there are many strategies for estimating outcomes for those with missing data (Polit, 2010).

Example of intention-to-treat analysis: Skrutkowski and colleagues (2008) used an RCT design to test the impact of a pivot nurse in oncology on symptom relief in patients with lung or breast cancer. They used an intention-to-treat analysis, even though participant loss over the course of the study was fairly high (31%). They stated that, "All participants' data were included, whether or not they provided survey data at each assessment period or died before completing the study" (p. 952).

In a crossover design, history is a potential threat both because an external event could differentially affect people in different treatment orderings and because the different orderings are in themselves a kind of differential history. *Substantive* analyses of the data involve comparing outcomes under treatment A versus treatment B. The analysis of bias, by contrast, involves comparing participants in the different orderings (e.g., A then B versus B then A). Significant differences between the two orderings is evidence of an **ordering bias**.

In summary, efforts to enhance the internal validity of a study should not end once the design strategy has been put in place. Researchers should seek additional opportunities to understand (and possibly to correct) the various threats to internal validity that can arise.

CONSTRUCT VALIDITY

Researchers conduct a study with specific exemplars of treatments, outcomes, settings, and people, which are stand-ins for broad constructs. Construct validity involves inferences from study particulars to the higher-order constructs that they are intended to represent. Construct validity is important because constructs are the means for linking the operations used in a study to a relevant conceptualization and to mechanisms for translating the resulting evidence into practice. If studies contain construct errors, there is a risk that the evidence will be misleading.

Enhancing Construct Validity

The first step in fostering construct validity is a careful explication of the treatment, outcomes, setting, and population constructs of interest; the next step is to carefully select instances that match those constructs as closely as possible. Construct validity is further cultivated when researchers assess the match between the exemplars and the constructs and the degree to which any "slippage" occurred.

Construct validity has most often been a concern to researchers in connection with the measurement of outcomes, an issue we discuss in Chapter 14. There is a growing interest, however, in the careful conceptualization and development of theory-based interventions in which the treatment itself has strong construct validity (see Chapter 26). It is just as important for the independent variable (whether it be an intervention or something not amenable to experimental manipulation) to be a strong instance of the construct of interest as it is for the measurement of the dependent variable to have strong correspondence to the outcome construct. In nonexperimental research, researchers do not create and manipulate the hypothesized cause, so ensuring construct validity of the independent variable is often more difficult.

Shadish and colleagues (2002) broadened the concept of construct validity to cover persons and settings as well as outcomes and treatments. For example, some nursing interventions specifically target groups that are characterized as "disadvantaged," but there is not always agreement on how this term is defined and operationalized. Researchers select specific people to represent the construct of a disadvantaged group about which inferences will be made, so it is important that the specific people are good exemplars of the underlying construct. The construct "disadvantaged" must be carefully delineated before a sample is selected. Similarly, if a researcher is interested in such settings as "immigrant neighborhoods" or "school-based clinics," these are constructs that require careful description—and the selection of exemplars that match those setting constructs. Qualitative description is often a powerful means of enhancing the construct validity of settings.

Threats to Construct Validity

Threats to construct validity are reasons that inferences from a particular study exemplar to an abstract

construct could be erroneous. Such a threat could occur if the operationalization of the construct fails to incorporate all the relevant characteristics of the underlying construct, or it could occur if it includes extraneous content—both of which are instances of a mismatch. Shadish and colleagues (2002) identified 14 threats to construct validity (their Table 3.1) and several additional threats specific to case-control designs (their Table 4.3). Among the most noteworthy threats are the following:

1. *Reactivity to the study situation*. As discussed in Chapter 9, participants may behave in a particular manner because they are aware of their role in a study (the Hawthorne effect). When people's responses reflect, in part, their perceptions of participation in research, those perceptions become part of the treatment construct under study. There are several ways to reduce this problem, including blinding, using outcome measures not susceptible to reactivity (e.g., data from hospital records), and using preintervention strategies to satisfy participants' desire to look competent or please the researcher.

Example of a possible Hawthorne effect:
Yap and colleagues (2009) evaluated the effect of tailored email messages on physical activity in manufacturing workers, using a two-group quasi-experimental design. Participants in *both* groups increased their activity, although increases were greater in the intervention group. The researchers speculated that the comparison group's improvement was probably a Hawthorne effect.

2. *Researcher expectancies*. A similar threat stems from the researcher's influence on participant responses through subtle (or not-so-subtle) communication about desired outcomes. When this happens, the researcher's expectations become part of the treatment (or nonmanipulated independent variable) construct that is being tested. Blinding is a strategy to reduce this threat, but another strategy is to use observations during the course of the study to detect verbal or behavioral signals of expectations and correct them.
3. *Novelty effects*. When a treatment is new, participants and research agents alike might alter their behavior. People may be either enthusiastic or skeptical about new methods of doing things. Results may reflect reactions to the novelty rather than to the intrinsic nature of an intervention, so the intervention construct is clouded by novelty content.
4. *Compensatory effects.* In intervention studies, *compensatory equalization* can occur if healthcare staff or family members try to compensate for the control group members' failure to receive a perceived beneficial treatment. The compensatory goods or services must then be part of the construct description of the treatment conditions. *Compensatory rivalry* is a related threat arising from the control group members' desire to demonstrate that they can do as well as those receiving a special treatment.
5. *Treatment diffusion or contamination.* Sometimes alternative treatment conditions can get blurred, which can impede good construct descriptions of the independent variable. This may occur when participants in a control group condition receive services similar to those available in the treatment condition. More often, however, blurring occurs when those in a treatment condition essentially put themselves into the control group by dropping out of the intervention. This threat can also occur in nonexperimental studies. For example, in case-control comparisons of smokers and nonsmokers, care must be taken during screening to ensure that study participants are, in fact, appropriately categorized (e.g., some people may consider themselves nonsmokers even though they smoke regularly, but only on weekends).

Construct validity requires careful attention to what we *call* things (i.e., construct labels) so that appropriate construct inferences can be made. Enhancing construct validity in a study requires careful thought before a study is undertaken, in terms of a well-considered explication of constructs, and also requires poststudy scrutiny to

assess the degree to which a match between operations and constructs was achieved.

EXTERNAL VALIDITY

External validity concerns the extent to which it can be inferred that relationships observed in a study hold true over variations in people, conditions, and settings, as well as over variations in treatments and outcomes. External validity has emerged as a very major concern in an EBP world in which there is an interest in generalizing evidence from tightly controlled research settings to real-world clinical practice settings.

External validity questions may take on several different forms (Shadish et al., 2002). We may wish to ask whether relationships observed with a study sample can be generalized to a larger population—for example, whether results from a smoking cessation program found effective with pregnant teenagers in Boston can be generalized to pregnant teenagers throughout the United States. Many EBP questions, however, are about going from a broad study group to a *particular* client—for example, whether the pelvic muscle exercises found to be effective in alleviating urinary incontinence in one study are an effective strategy for Linda Smith. Other external validity questions are about generalizing to types of people, settings, situations, or treatments unlike those in the research (Polit & Beck, 2010). For example, can findings about a pain-reduction treatment in a study of Australian women be generalized to men and women in Canada? Or, would a 6-week intervention to promote dietary changes in patients with diabetes be equally effective if the content were condensed into a 3-week program? Sometimes new studies are needed to answer questions about external validity, but sometimes external validity can be enhanced by decisions that the researcher makes in designing a study.

Enhancements to External Validity

One aspect of external validity concerns the *representativeness* of the exemplars used in the study. For example, if the sample is selected to be representative of a population to which the researcher wishes to generalize the results, then the findings can more readily be applied to that population (see Chapter 12 for sampling designs). Similarly, if the settings in which the study occurs are representative of the clinical settings in which the findings might be applied, then inferences about relevance in those other settings can be strengthened.

An important concept for external validity is *replication.* Multisite studies are powerful because more confidence in the generalizability of the results can be attained if results have been replicated in several sites—particularly if the sites are different on dimensions considered important (e.g., size, nursing skill mix, and so on). Studies with a varied sample of participants can test whether study results are replicated for subgroups of the sample—for example, whether benefits from an intervention apply to men *and* women, or older *and* younger patients. Systematic reviews are a crucial aid to external validity precisely because they assess relationships in replicated studies across time, space, people, and settings.

Another issue concerns attempts to use or create study situations as similar as possible to real-world circumstances. The real world is a "messy" place, lacking the standardization imposed in studies. Yet, external validity can be jeopardized if study conditions are too artificial. For example, if nurses require 5 days of training to implement a promising intervention, we might ask how realistic it would be for administrators to devote resources to such an intervention.

Threats to External Validity

In the previous chapter, we discussed *interaction effects* that can occur in a factorial design when two treatments are simultaneously manipulated. The interaction question is whether the effects of treatment A hold (are comparable) for all levels of treatment B. Conceptually, questions regarding external validity are similar to this interaction question. Threats to external validity concern ways in which relationships between variables might interact with

or be moderated by variations in people, settings, time, and conditions. Shadish and colleagues (2002) described several threats to external validity, such as the following two:

1. *Interaction between relationship and people*. An effect observed with certain types of people might not be observed with other types of people. A common complaint about some RCTs is that many people are excluded not because they would not benefit from the treatment, but rather because they cannot provide needed research data (e.g., cognitively impaired patients, non-English speakers). During the 1980s, the widely held perception that many clinical trials were conducted primarily with white males led to policy changes to ensure that treatment by gender and ethnicity subgroup interactions were explored.
2. *Interaction between causal effects and treatment variation.* An innovative treatment might be effective because it is paired with other elements, and sometimes those elements are intangible—for example, an enthusiastic and dedicated project director. The same "treatment" could never be fully replicated, and thus different results could be obtained in subsequent tests.

Shadish and colleagues (2002) noted that moderators of relationships are the norm, not the exception. With interventions, for example, it is normal for a treatment to "work better" for some people than for others. Thus, in thinking about external validity, the primary issue is whether there is constancy of a relationship (or constancy of causation), and not whether the *magnitude* of the effect is constant.

TRADE-OFFS AND PRIORITIES IN STUDY VALIDITY

Quantitative researchers strive to design studies that are strong with respect to all four types of study validity. Sometimes, efforts to increase one type of validity will also benefit another type. In some instances, however, the requirements for ensuring one type of validity interfere with the possibility of achieving others.

For example, suppose we went to great lengths to ensure intervention fidelity in an RCT. Our efforts might include strong training of staff, careful monitoring of intervention delivery, manipulation checks, and steps to maximize participants' adherence to treatment. Such efforts would have positive effects on statistical conclusion validity because the treatment was made as powerful as possible. Internal validity would be enhanced if attrition biases were minimized as a result of high adherence. Intervention fidelity would also improve the construct validity of the treatment because the content delivered and received would better match the underlying construct. But what about external validity? All of the actions undertaken to ensure that the intervention is strong, construct-valid, and administered according to plan are not consistent with the realities of clinical settings. People are not normally paid to adhere to treatments, nurses are not monitored and corrected to ensure that they are following a script, training in the use of new protocols is usually brief, and so on.

This example illustrates that researchers need to give careful thought to how design decisions may affect various aspects of study validity. Of particular concern are trade-offs between internal and external validity.

Internal Validity and External Validity

Tension between the goals of achieving internal validity and external validity is pervasive. Many control mechanisms that are designed to rule out competing explanations for hypothesized cause-and-effect relationships make it difficult to infer that the relationship holds true in uncontrolled real-life settings.

Internal validity was long considered the "*sine qua non*" of experimental research (Campbell & Stanley, 1963). The rationale was this: If there is insufficient evidence that an intervention really caused an effect, why worry about generalizing the results? This high priority given to internal validity,

however, is somewhat at odds with the current emphasis on evidence-based practice. A question that some are now posing is this: If study results can't be generalized to real-world clinical settings, who *cares* if the study has strong internal validity? Clearly, both internal and external validity are important to building an evidence base for nursing practice.

There are several "solutions" to the conflict between internal and external validity. The first (and perhaps most prevalent) approach is to emphasize one and sacrifice the other. Following a long tradition of field experimentation based on Campbell and Stanley's advice, it is often external validity that is sacrificed.

A second approach in some medical trials is to use a phased series of studies. In the earlier phase, there are tight controls, strict intervention protocols, and stringent criteria for including people in the RCT. Such studies are **efficacy studies**. Once the intervention has been deemed to be effective under tightly controlled conditions in which internal validity was the priority, it is tested with larger samples in multiple sites under less restrictive conditions, in **effectiveness studies** that emphasize external validity.

A third approach is to compromise. There has been recent interest in promoting designs that aim to achieve a balance between internal and external validity in a single intervention study. We discuss such *practical* (or *pragmatic*) *clinical trials* in Chapter 11.

Efforts to improve the generalizability of healthcare research evidence have given rise to a framework for designing and evaluating intervention research called the **RE-AIM framework** (Glasgow, 2006). The framework involves a scrutiny of five aspects of a study: its **R**each, **E**fficacy, **A**doption, **I**mplementation, and **M**aintenance. *Reach* means reaching the intended population of potential beneficiaries, which concerns the extent to which study participants have characteristics that reflect those of that population. *Efficacy* concerns intervention impacts on critical outcomes. *Adoption* concerns the number and representativeness of settings and staff who are willing to implement the intervention. *Implementation* concerns the consistency of delivering the intervention as intended, and also intervention costs. The last component, *maintenance*, involves a consideration of the extent to which, at the individual level, outcomes are maintained over time and, at the institutional level, the intervention becomes part of routine practices and policies. Table 10.3 summarizes some key planning questions for each of these five components. Detailed information about this new framework and advice on how to enhance and assess the five components is available at *www.re-aim.org*.

Example of a study using RE-AIM: Whittemore and colleagues (2009) used the RE-AIM model as the organizing framework for their pilot study of a diabetes prevention program in primary care settings. The study appears in its entirety in Appendix D of the accompanying *Resource Manual*.

TIP: The Toolkit section of the *Resource Manual* includes a table listing a number of strategies that can be used to enhance the external validity of a study. The table identifies the potential consequence of each strategy for other types of study validity.

Prioritization and Design Decisions

Unfortunately, it is impossible to avoid all possible threats to study validity. By understanding the various threats, however, you can come to conclusions about the kinds of trade-offs you are willing to make to achieve study goals. Some threats are more worrisome than others in terms of both likelihood of occurrence and consequences to the inferences you would like to make. And some threats are more costly to avoid than others. Resources available for a study must be allocated so that there is a correspondence between expenditures and the importance of different types of validity. For example, with a fixed budget, you need to decide whether it is better to increase the size of the sample and hence power (statistical conclusion validity), or to use the money on efforts to reduce attrition (internal validity).

TABLE 10.3 Key Planning Questions within the RE-AIM Framework

RE-AIM COMPONENT	PLANNING QUESTIONS
Reach	• How can I reach those who need the intervention? • How can I design the intervention and the research so as to persuade those who need it to try it?
Efficacy	• How can I plan the intervention to maximize its efficacy? • How can I design the research to maximize the potential to detect its effects?
Adoption	• How can I best select study sites to represent environments where the intervention might be implemented? • How can I develop organizational support for the delivery of my intervention?
Implementation	• What can I do to enhance the likelihood that the intervention is delivered properly? • How can I best assess and document the extent to which intervention fidelity occurred?
Maintenance	• How can I design the intervention so as to encourage long-term maintenance of needed behaviors? • What can I do to enhance the likelihood that the intervention is maintained and delivered over the long term?

The point here is that you should make conscious decisions about how to structure a study to address validity concerns. Every design decision has both a "payoff" and a cost in terms of study integrity. Being cognizant of the effects that design decisions have on the quality of research evidence is a responsibility that nurse researchers should attend to so that their evidence can have the largest possible impact on clinical practice.

TIP: A useful strategy is to create a matrix that lists various design decisions in the first column (e.g., randomization, crossover design), and then use the next four columns to identify the potential impact of those options on the four types of study validity. (In some cells, there may be no entry if there are no consequences of a design element for a given type of validity). A sixth column could be added for estimates of the design element's financial implications, if any. The Toolkit section of the accompanying *Resource Manual* includes a model matrix as a Word document for you to use and adapt.

CRITIQUING GUIDELINES FOR STUDY VALIDITY

In critiquing a research report to evaluate its potential to contribute to nursing practice, it is crucial to make judgments about the extent to which threats to validity were minimized—or, at least, assessed and taken into consideration during the interpretation of the results. The guidelines in Box 10.1 focus on validity-related issues to further help you in the critique of quantitative research designs. Together with the critiquing guidelines in the previous chapter, they are likely to be the core of a strong critical evaluation of the evidence that quantitative studies yield. From an EBP perspective, it is important to remember that drawing inferences about causal relationships relies not only on how high up on the evidence hierarchy a study is (Figure 2.1), but also, for any given level of the hierarchy, how successful the researcher was in managing study validity and balancing competing validity demands.

BOX 10.1 Guidelines for Critiquing Design Elements and Study Validity in Quantitative Studies

1. Was there adequate statistical power? Did the manner in which the independent variable was defined and operationalized create strong contrasts that enhanced statistical power? Was precision enhanced by controlling confounding variables? If hypotheses were not supported (e.g., a hypothesized relationship was not found), is it possible that statistical conclusion validity was compromised?
2. In intervention studies, is there evidence that attention was paid to intervention fidelity? For example, were staff adequately trained? Was the implementation of the intervention monitored? Was attention paid to both the delivery and receipt of the intervention?
3. What evidence does the report provide that selection biases were eliminated or minimized? What steps were taken to control confounding participant characteristics that could affect the equivalence of groups being compared? Were these steps adequate?
4. To what extent did the study design rule out the plausibility of other threats to internal validity, such as history, attrition, maturation, and so on? What are your overall conclusions about the internal validity of the study?
5. Were there any major threats to the construct validity of the study? In intervention studies, was there a good match between the underlying conceptualization of the intervention and its operationalization? Was the intervention "pure" or was it confounded with extraneous content, such as researcher expectations? Was the setting or site a good exemplar of the type of setting envisioned in the conceptualization?
6. Was the context of the study sufficiently described to enhance its capacity for external validity? Were the settings or participants representative of the types to which results were designed to be generalized?
7. Overall, did the researcher appropriately balance validity concerns? Was attention paid to certain types of threats (e.g., internal validity) at the expense of others (e.g., external validity)?

RESEARCH EXAMPLE

We conclude this chapter with an example of a study that demonstrated careful attention to many aspects of study validity.

Study: Effects of abdominal massage in management of constipation—A randomized controlled trial (Lämås et al., 2009)

Statement of Purpose: The purpose of the study was to assess the effect of an abdominal massage on gastrointestinal functions and use of laxatives in people with constipation.

Treatment Groups: There were two treatment groups: an intervention group that received an abdominal massage 5 days per week for 8 weeks in addition to previously prescribed laxatives, and a control group that continued with usual laxatives and treatments but no massage.

Method: A sample of 60 people with constipation was recruited from a Swedish community via local newspapers and notices at care centers. Eligible participants were randomly assigned to treatment groups by block randomization, with four patients per block. Gastrointestinal function was assessed with a standardized instrument at baseline, 4 weeks, and 8 weeks. Participants also maintained a daily diary in which they recorded information about bowel movements and use of remedies such as laxatives and fiber.

Additional Study Validity Efforts: The researchers estimated how large a sample was needed to achieve adequate power for statistical conclusion validity, using a procedure called power analysis (Chapter 12). Study protocols and a manual were developed to standardize the massage intervention. Massage interventionists were trained by the lead author. Data were gathered by self-administration (the data collectors were not blinded). Selection bias was assessed by comparing the baseline characteristics of the two groups, who were comparable with regard to demographic characteristics (e.g., age, sex), laxative use, and most indexes of gastrointestinal function. However, those in the intervention group had higher constipation scores, so these baseline scores were statistically adjusted in estimating

intervention effects 8 weeks later. Attrition was similar in both groups (10% per group). An intention-to-treat analysis was performed by estimating missing outcome values for those who dropped out of the study.

Key Findings: Those in the intervention group had significantly better outcomes at 8 weeks than those on the control group with regard to constipation and abdominal pain. The massage group also had significantly more bowel movements. The groups had similar usage of laxatives at the end of the study, suggesting massage might be an effective complement to, but not substitute for, laxatives in this population.

SUMMARY POINTS

- Study validity concerns the extent to which appropriate inferences can be made. **Threats to validity** are reasons that an inference could be wrong. A key function of quantitative research design is to rule out validity threats by exercising various types of control.
- Control over confounding participant characteristics is key to managing many validity threats. The best control method is randomization to treatment conditions, which effectively controls all confounding variables—especially within the context of a crossover design.
- When randomization is not possible, other control methods include **homogeneity (**the use of a homogeneous sample to eliminate variability on confounding characteristics); blocking or stratifying, as in the case of a randomized block design; **pair matching** participants on key variables to make groups more comparable (or **balancing** groups to achieve comparability); and **statistical control** to remove the effect of a confounding variable statistically (e.g., through **analysis of covariance**).
- Homogeneity, stratifying, matching, and statistical control share two disadvantages: Researchers must know in advance which variables to control, and they can rarely control all of them.
- Four types of validity affect the rigor of a quantitative study: statistical conclusion validity, internal validity, construct validity, and external validity.
- **Statistical conclusion validity** concerns the validity of inferences that there is an empirical relationship between variables (most often, the presumed cause and the effect).
- Threats to statistical conclusion validity include low **statistical power** (the ability to detect true relationships among variables), low **precision** (the exactness of the relationships revealed after controlling confounding variables), and factors that undermine a strong operationalization of the independent variable (e.g., a treatment).
- **Intervention** (or **treatment) fidelity** concerns the extent to which the implementation of a treatment is faithful to its plan. Intervention fidelity is enhanced through standardized treatment protocols, careful training of intervention agents, monitoring of the delivery and receipt of the intervention, **manipulation checks**, and steps to promote **treatment adherence** and avoid **contamination of treatments**.
- **Internal validity** concerns inferences that outcomes were caused by the independent variable, rather than by factors extraneous to the research. Threats to internal validity include temporal ambiguity (lack of clarity about whether the presumed cause preceded the outcome), **selection** (preexisting group differences), **history** (the occurrence of events external to an independent variable that could affect outcomes), **maturation** (changes resulting from the passage of time), **mortality** (effects attributable to attrition), **testing** (effects of a pretest), and **instrumentation** (changes in the way data are gathered).
- Internal validity can be enhanced through judicious design decisions, but can also be addressed analytically (e.g., through an analysis of selection or attrition biases). When people withdraw from a study, an **intention-to-treat analysis** (analyzing outcomes for all people in their original treatment conditions) is preferred to a **per-protocol analysis** (analyzing outcomes only for those who received the full treatment as assigned) for maintaining the integrity of randomization.
- **Construct validity** concerns inferences from the particular exemplars of a study (e.g., the specific treatments, outcomes, people, and settings) to the

higher-order constructs that they are intended to represent. The first step in fostering construct validity is a careful explication of those constructs.

- Threats to construct validity can occur if the operationalization of a construct fails to incorporate all of the relevant characteristics of the construct or if it includes extraneous content. Examples of such threats include *subject reactivity*, *researcher expectancies*, *novelty effects*, *compensatory effects*, and *treatment diffusion*.
- **External validity** concerns inferences about the extent to which study results can be generalized—that is, about whether relationships observed in a study hold true over variations in people, settings, outcome measures, and treatments. External validity can be enhanced by selecting *representative* people, settings, and so on and through replication.
- Researchers need to prioritize and recognize trade-offs among the various types of validity, which sometimes compete with each other. Tensions between internal and external validity are especially prominent. One solution has been to begin with a study that emphasizes internal validity (**efficacy studies**) and then if a causal relationship can be inferred, to undertake **effectiveness studies** that emphasize external validity.
- The **RE-AIM framework** (*Reach*, *Efficacy*, *Adoption*, *Implementation*, and *Maintenance*) is a model for designing and evaluating intervention research that is strong on multiple forms of study validity.

STUDY ACTIVITIES

Chapter 10 of the *Study Guide for Nursing Research*: *Generating and Assessing Evidence for Nursing Practice*, *9th edition*, offers exercises and study suggestions for reinforcing concepts presented in this chapter. In addition, the following study questions can be addressed:

1. How do you suppose the use of identical twins in a study could enhance control?
2. To the extent possible, apply the questions in Box 10.1 to the massage intervention study described at the end of the chapter (Lämås, et al., 2009).

STUDIES CITED IN CHAPTER 10

Lämås, K., Lindholm, L., Stenlund, H., Engstrom, B., & Jacobsson, C. (2009). Effects of abdominal massage in management of constipation—A randomized controlled trial. *International Journal of Nursing Studies*, *46*, 759–767.

Lee, L. Y., Lee, D. T., & Woo, J. (2009). Tai Chi and health-related quality of life in nursing home residents. *Journal of Nursing Scholarship*, *41*, 35–43.

Luttik, M., Jaarsma, T., Lesman, I., Sanderman, R., & Hagedoorn, M. (2009). Quality of life in partners of people with congestive heart failure. *Journal of Advanced Nursing*, *65*, 1442–1451.

Ngai, F., Chan, S., & Ip, W. (2010). Predictors and correlates of maternal role competence and satisfaction. *Nursing Research*, *59*, 185–193.

Radziewicz, R., Rose, J., Bowman, K., Berila, R., O'Toole, E., & Given, B. (2009). Establishing treatment fidelity in a coping and communication support telephone intervention for aging patients with advanced cancer and their family caregivers. *Cancer Nursing*, *32*, 193–203.

Resnick, B., Luisi, D., & Vogel, A. (2008). Testing the Senior Exercise Self-efficacy Project (SESEP) for use with urban dwelling minority older adults. *Public Health Nursing*, *25*, 221–234.

Skrutkowski, M., Saucier, A., Eades, M., Swidzinski, M., Ritchie, J., Marchionni, C., Ladouceur, M. (2008). Impact of a pivot nurse in oncology on patients with lung or breast cancer. *Oncology Nursing Forum*, *35*, 948–954.

Whittemore, R., Melkus, G., Wagner, G., Dziura, J., Northrup, V., & Grey, M. (2009). Translating the diabetes prevention program to primary care. *Nursing Research*, *58*, 2–12.

Yap, T., Davis, L., Gates, D., Hemmings, A., & Pan, W. (2009). The effect of tailored emails in the workplace. *AAOHN Journal*, *57*, 267–273.

Methodologic and nonresearch references cited in this chapter can be found in a separate section at the end of the book.

12 Sampling in Quantitative Research

Sampling is familiar to us all. In the course of daily activities, we make decisions and draw conclusions through sampling. A nursing student may select an elective course by sampling two or three classes on the first day of the semester. Patients may generalize about nursing care in a hospital based on the care they received from a sample of nurses. We all come to conclusions about phenomena based on exposure to a limited portion of those phenomena.

Researchers, too, obtain data from samples. In testing the efficacy of a new asthma medication, researchers reach conclusions without giving the drug to all asthmatic patients. Researchers, however, cannot afford to draw conclusions about intervention effects or inter-relationships among variables based on a sample of only three or four people. The consequences of making faulty decisions are more momentous in research than in private decision making.

Quantitative researchers seek to select samples that will allow them to achieve statistical conclusion validity and to generalize their results. They develop a **sampling plan** that specifies in advance how participants are to be selected and how many to include. Qualitative researchers, by contrast, make sampling decisions during the course of data collection, and typically do not have a formal sampling plan. This chapter discusses sampling issues for quantitative studies. Sampling for qualitative research is discussed in Chapter 21.

BASIC SAMPLING CONCEPTS

Let us begin by considering some terms associated with sampling—terms that are used primarily (but not exclusively) in quantitative research.

Populations

A **population** is the entire aggregation of cases in which a researcher is interested. For instance, if we were studying American nurses with doctoral degrees, the population could be defined as all U.S. citizens who are registered nurses (RNs) and who have a PhD, DNSc, DNP, or other doctoral-level degree. Other possible populations might be all male patients who had cardiac surgery in St. Peter's Hospital in 2010, all women with irritable bowel syndrome in Sydney, or all children in Canada with cystic fibrosis. As this list illustrates, a population may be broadly defined to involve thousands of people, or narrowly specified to include only hundreds.

Populations are not restricted to humans. A population might consist of all hospital records in a particular hospital or all blood samples at a particular

laboratory. Whatever the basic unit, the population comprises the aggregate of elements in which the researcher is interested.

It is useful to make a distinction between target and accessible populations. The **accessible population** is the aggregate of cases that conform to designated criteria *and* that are accessible for a study. The **target population** is the aggregate of cases about which the researcher would like to generalize. A target population might consist of all diabetic people in the United States, but the accessible population might consist of all diabetic people who attend a particular clinic. Researchers usually sample from an accessible population and hope to generalize to a target population.

TIP: A key issue for evidence-based practice is information about the populations on whom research has been conducted. Many quantitative researchers fail to identify their target population, or to discuss the generalizability of the results. The population of interest needs to be carefully considered in planning and reporting a study.

Eligibility Criteria

Researchers must specify criteria that define who is in the population. Consider the population, American nursing students. Does this population include students in all types of nursing programs? How about RNs returning to school for a bachelor's degree? Or students who took a leave of absence for a semester? Do foreign students enrolled in American nursing programs qualify? Insofar as possible, the researcher must consider the exact criteria by which it could be decided whether an individual would or would not be classified as a member of the population. The criteria that specify population characteristics are the **eligibility criteria** or **inclusion criteria**. Sometimes, a population is also defined in terms of characteristics that people must *not* possess (i.e., the **exclusion criteria**). For example, the population may be defined to exclude people who cannot speak English.

Specifications about the population should be driven, to the extent possible, by theoretical considerations. In thinking about ways to define the population and delineate eligibility criteria, it is important to consider whether the resulting sample is likely to be a good exemplar of the population construct in which you are interested. A study's construct validity is enhanced when there is a good match between the eligibility criteria and the population construct.

Of course, inclusion or exclusion criteria for a study often reflect considerations other than substantive concerns. Eligibility criteria may reflect one or more of the following:

- *Costs*. Some criteria reflect cost constraints. For example, when non–English-speaking people are excluded, this does not usually mean that researchers are uninterested in non–English speakers, but rather that they cannot afford to hire translators and multilingual data collectors.
- *Practical constraints*. Sometimes, there are other practical constraints, such as difficulty including people from rural areas, people who are hearing impaired, and so on.
- *People's ability to participate in a study*. The health condition of some people may preclude their participation. For example, people with mental impairments, who are in a coma, or who are in an unstable medical condition may need to be excluded.
- *Design considerations*. As noted in Chapter 10, it is sometimes advantageous to a study's internal validity to define a homogeneous population as a means of controlling confounding variables.

The criteria used to define a population for a study have implications for the interpretation of the results and, of course, the external validity of the findings.

Example of inclusion and exclusion criteria: Hafsteindóttir and colleagues (2010) studied malnutrition in hospitalized neurologic patients. Study participants had to be diagnosed with a neurologic or neurosurgical disease and speak Dutch. Patients were excluded if they were bed-bound and if their health condition made participation impossible.

Samples and Sampling

Sampling is the process of selecting cases to represent an entire population so that inferences about the population can be made. A **sample** is a subset of population **elements**, which are the most basic units about which data are collected. In nursing research, elements are usually humans.

Samples and sampling plans vary in quality. *Two key considerations in assessing a sample in a quantitative study are its representativeness and size*. A **representative sample** is one whose key characteristics closely approximate those of the population. If the population in a study of blood donors is 50% male and 50% female, then a representative sample would have a similar gender distribution. If the sample is not representative of the population, the study's external validity (and construct validity) is at risk.

Unfortunately, there is no way to make sure that a sample is representative without obtaining information from the population. Certain sampling procedures are less likely to result in biased samples than others, but a representative sample can never be guaranteed. Researchers operate under conditions in which error is possible. Quantitative researchers strive to minimize errors and, when possible, to estimate their magnitude.

Sampling designs are classified as either probability sampling or nonprobability sampling. **Probability sampling** involves random selection of elements. In probability sampling, researchers can specify the probability that an element of the population will be included in the sample. Greater confidence can be placed in the representativeness of probability samples. In **nonprobability samples**, elements are selected by nonrandom methods. There is no way to estimate the probability that each element has of being included in a nonprobability sample, and every element usually does *not* have a chance for inclusion.

Strata

Sometimes, it is useful to think of populations as consisting of subpopulations, or **strata**. A stratum is a mutually exclusive segment of a population, defined by one or more characteristics. For instance, suppose our population was all RNs in the United States. This population could be divided into two strata based on gender. Or, we could specify three strata of nurses younger than 30 years of age, nurses aged 30 to 45 years, and nurses 46 years or older. Strata are often used in sample selection to enhance the sample's representativeness.

Staged Sampling

Samples are sometimes selected in multiple stages, in what is called **multistage sampling**. In the first stage, large units (such as hospitals or nursing homes) are selected. Then, in a later stage, individual people are sampled. In staged sampling, it is possible to combine probability and nonprobability sampling. For example, the first stage can involve the deliberate (nonrandom) selection of study sites. Then, people within the selected sites can be selected through random procedures.

Sampling Bias

Researchers work with samples rather than with populations because it is cost-effective to do so. Researchers typically do not have the resources to study all members of a population.

It is often possible to obtain reasonably accurate information from a sample, but data from samples *can* lead to erroneous conclusions. Finding 100 people willing to participate in a study is seldom difficult. It is considerably harder to select 100 people who are not a biased subset of the population. **Sampling bias** refers to the systematic over-representation or under-representation of a population segment on a characteristic relevant to the research question.

As an example of consciously biased selection, suppose we were investigating patients' responsiveness to nurses' touch and decide to recruit the first 50 patients meeting eligibility criteria. We decide, however, to omit Mr. Z from the sample because he has been hostile to nursing staff. Mrs. X, who has

just lost a spouse, is also bypassed because she is under stress. We have made conscious decisions to exclude certain people, and the decisions do not reflect bona fide eligibility criteria. This can lead to bias because responsiveness to nurses' touch (the dependent variable) may be affected by patients' feelings about nurses or their emotional state.

Sampling bias often occurs unconsciously, however. If we were studying nursing students and systematically interviewed every 10th student who entered the nursing school library, the sample would be biased in favor of library-goers, even if we were conscientious about including every 10th student regardless of his or her age, gender, or other traits.

TIP: Internet surveys are attractive because they can be distributed to people all over the world. However, there is an inherent bias in such surveys, unless the population is defined as people who have easy access to, and comfort with, a computer and the Internet.

Sampling bias is partly a function of population homogeneity. If population elements were all identical with respect to key attributes, then any sample would be as good as any other. Indeed, if the population were completely homogeneous, that is, exhibited no variability at all, then a *single* element would be sufficient to draw conclusions about the population. For many physiologic attributes, it may be safe to assume high homogeneity. For example, the blood in a person's veins is relatively homogeneous and so a single blood sample is adequate. For most human attributes, however, homogeneity is the exception rather than the rule. Age, health status, stress, motivation—all these attributes reflect human heterogeneity. When variation occurs in the population, then similar variation should be reflected, to the extent possible, in a sample.

TIP: One easy way to increase a study's generalizability is to select participants from multiple sites (e.g., from different hospitals, nursing homes, communities, etc.). Ideally, the different sites would be sufficiently divergent that good representation of the population would be obtained.

NONPROBABILITY SAMPLING

Nonprobability sampling is less likely than probability sampling to produce representative samples. Despite this fact, most studies in nursing and other disciplines rely on nonprobability samples. Four types of nonprobability sampling in quantitative studies are convenience, quota, consecutive, and purposive.

Convenience Sampling

Convenience sampling entails using the most conveniently available people as participants. A faculty member who distributes questionnaires to nursing students in a class is using a convenience sample. The nurse who conducts a study of teenage risk taking at a local high school is also relying on a convenience sample. The problem with convenience sampling is that those who are available might be atypical of the population with regard to critical variables.

Convenience samples do not necessarily comprise individuals known to the researchers. Stopping people at a street corner to conduct an interview is sampling by convenience. Sometimes, researchers seeking people with certain characteristics place an advertisement in a newspaper, put up signs in clinics, or post messages in chat rooms on the Internet. These approaches are subject to bias because people select themselves as pedestrians on certain streets or as volunteers in response to posted notices.

Snowball sampling (also called *network sampling* or *chain sampling*) is a variant of convenience sampling. With this approach, early sample members (called **seeds**) are asked to refer other people who meet the eligibility criteria. This sampling method is often used when the population is people with characteristics who might otherwise be difficult to identify (e.g., people who are afraid of hospitals). Snowballing begins with a few eligible participants and then continues on the basis of participant referrals.

Convenience sampling is the weakest form of sampling. In heterogeneous populations, there is no other sampling approach in which the risk of

TABLE 12.1 Numbers and Percentages of Students in Strata of a Population, Convenience Sample, and Quota Sample

STRATA	POPULATION	CONVENIENCE SAMPLE	QUOTA SAMPLE
Male	100 (20%)	5 (5%)	20 (20%)
Female	400 (80%)	95 (95%)	80 (80%)
Total	500 (100%)	100 (100%)	100 (100%)

sampling bias is greater. Yet, convenience sampling is the most commonly used method in many disciplines.

Example of a convenience sample: Peddle and colleagues (2009) studied factors that correlated with adherence to supervised exercise in patients awaiting surgery for suspected malignant lung lesions. Their sample of patients was described as a sample of convenience.

TIP: Rigorous methods of sampling *hidden populations,* such as the homeless or injection drug users, are emerging. Because standard probability sampling is inappropriate for such hidden populations, a method called **respondent-driven sampling (RDS)**, a variant of snowball sampling, has been developed. RDS, unlike traditional snowballing, allows the assessment of relative inclusion probabilities based on mathematical models (Magnani et al., 2005).

Quota Sampling

A **quota sample** is one in which the researcher identifies population strata and determines how many participants are needed from each stratum. By using information about population characteristics, researchers can ensure that diverse segments are represented in the sample, preferably in the proportion in which they occur in the population.

Suppose we were interested in studying nursing students' attitude toward working with AIDS patients. The accessible population is a school of nursing with 500 undergraduate students; a sample of 100 students is desired. The easiest procedure would be to distribute questionnaires in classrooms through convenience sampling. We suspect, however, that male and female students have different attitudes, and a convenience sample might result in too many men or women. Table 12.1 presents fictitious data showing the gender distribution for the population and for a convenience sample (second and third columns). In this example, the convenience sample over-represents women and under-represents men. We can, however, establish "quotas" so that the sample includes the appropriate number of cases from both strata. The far-right column of Table 12.1 shows the number of men and women required for a quota sample for this example.

You may better appreciate the dangers of a biased sample with a concrete example. Suppose a key study question was, "Would you be willing to work on a unit that cared exclusively for AIDS patients?" The number and percentage of students in the population who would respond "yes" are shown in the first column of Table 12.2. We would not know these values—they are shown to illustrate a point. Within the population, men are more likely than women to say they would work on a unit with AIDS patients, yet men were under-represented in the convenience sample. As a result, population and sample values on the outcome are discrepant: Nearly twice as many students in the population are favorable toward working with AIDS patients (20%) than we would conclude based on results from the convenience sample (11%). The quota sample does a better

TABLE 12.2 Students Willing to Work on AIDS Unit, in the Population, Convenience Sample, and Quota Sample

	POPULATION	CONVENIENCE SAMPLE	QUOTA SAMPLE
Willing males (number)	28	2	6
Willing females (number)	72	9	13
Total number of willing students	100	11	19
Total number of all students	500	100	100
Percentage willing	20%	11%	19%

job of reflecting the views of the population (19%). In actual research situations, the distortions from a convenience sample may be smaller than in this example, but could be larger as well.

Quota sampling does not require sophisticated skills or a lot of effort. Many researchers who use a convenience sample could profitably use quota sampling. Stratification should be based on one or more variables that would reflect important differences in the dependent variable. Such variables as gender, ethnicity, education, and medical diagnosis may be good stratifying variables.

Procedurally, quota sampling is like convenience sampling. The people in any subgroup are a convenience sample from that stratum of the population. For example, the initial sample of 100 students in Table 12.1 constituted a convenience sample from the population of 500. In the quota sample, the 20 men constitute a convenience sample of the 100 men in the population. Because of this fact, quota sampling shares many of the same weaknesses as convenience sampling. For instance, if a researcher is required by a quota-sampling plan to interview 10 men between the ages of 65 and 80 years, a trip to a nursing home might be the most convenient method of obtaining participants. Yet this approach would fail to represent the many older men living independently in the community. Despite its limitations, quota sampling is a major improvement over convenience sampling.

Example of a quota sample: Fox and colleagues (2009) explored perceptions of bed days in patients receiving extended in-patient services for the management of chronic illness. The study used patients from a larger study that used quota sampling to ensure equal representation of people who had different levels of bed days. The strata were defined as people with 0, 2 to 4, and 5 to 7 bed days per week.

Consecutive Sampling

Consecutive sampling involves recruiting *all* of the people from an accessible population who meet the eligibility criteria over a specific time interval, or for a specified sample size. For example, in a study of ventilator-associated pneumonia in ICU patients, if the accessible population were patients in an ICU of a specific hospital, a consecutive sample might consist of all eligible patients admitted to that ICU over a 6-month period. Or it might be the first 250 eligible patients admitted to the ICU, if 250 were the targeted sample size.

Consecutive samples can be selected either for a retrospective or prospective time period. For example, the sample could include every patient who visited a diabetic clinic in the previous 30 days. Or, it could include all of the patients who will enroll in the clinic in the next 30 days.

Consecutive sampling is a far better approach than sampling by convenience, especially if the sampling period is sufficiently long to deal with

potential biases that reflect seasonal or other time-related fluctuations. When all members of an accessible population are invited to participate in a study over a fixed time period, the risk of bias is greatly reduced. Consecutive sampling is often the best possible choice when there is "rolling enrollment" into a contained accessible population.

Example of a consecutive sample: O'Meara and colleagues (2008) conducted a study to evaluate factors associated with interruptions in enteral nutrition delivery in mechanically ventilated critically ill patients. A consecutive sample of 59 ICU patients who required mechanical ventilation and were receiving enteral nutrition participated in the study.

Purposive Sampling

Purposive sampling or *judgmental sampling* uses researchers' knowledge about the population to select sample members. Researchers might decide purposely to select people who are judged to be typical of the population or particularly knowledgeable about the issues under study. Sampling in this subjective manner, however, provides no external, objective method for assessing the typicalness of the selected participants. Nevertheless, this method can be used to advantage in certain situations. Newly developed instruments can be effectively pretested and evaluated with a purposive sample of diverse types of people. Purposive sampling is often used when researchers want a sample of experts, as in the case of a needs assessment using the key informant approach or in Delphi surveys.

Purposive sampling is also a good approach in two-staged sampling. That is, sites can first be sampled purposively, and then people can be sampled in some other fashion, as in the following example:

Example of purposive sampling: Dudley-Brown and Freivogel (2009) field tested alternative intake tools for identifying patients at high risk for colorectal cancer in gastroenterology clinics. They began by purposively selecting six sites in four states. Their goal was to select sites so as to "approximate a representative sample for ethnicity and age" (p. 10). In the next stage of sampling, the researchers recruited a consecutive sample of patients over a 2-month period.

Evaluation of Nonprobability Sampling

Except for some consecutive samples, nonprobability samples are rarely representative of the population. When every element in the population does not have a chance of being included in the sample, it is likely that some segment of it will be systematically under-represented. When there is sampling bias, there is a chance that the results could be misleading, and efforts to generalize to a broader population could be misguided.

Nonprobability samples will continue to predominate, however, because of their practicality. Probability sampling requires skill and resources, so there may be no option but to use a nonprobability approach. Strict convenience sampling without explicit efforts to enhance representativeness, however, should be avoided. Indeed, it could be argued that quantitative researchers would do better at achieving representative samples for generalizing to a population if they had an approach that were more purposeful (Polit & Beck, 2010).

Quota sampling is a semi-purposive sampling strategy that is far superior to convenience sampling because it seeks to ensure sufficient representation within key strata of the population. Another purposive strategy for enhancing generalizability is deliberate multisite sampling. For instance, a convenience sample could be obtained from two communities known to differ socioeconomically so that the sample would reflect the experiences and views of both lower- and middle-class participants. In other words, if the population is known to be heterogeneous, you should take steps to capture important variation in the sample.

Even in one-site studies in which convenience sampling is used, researchers can (and should) make an effort to explicitly add cases to correspond more closely to population parameters. Kerlinger and Lee (2000) advised researchers to check their sample for easily verified expectations. For example, if half the population is known to be male, then the researcher can check to see if approximately half the sample is male and use outreach to recruit more males if necessary. Shadish and colleagues (2002) also argued for more purposive sampling,

noting that deliberate heterogeneous sampling on presumptively important dimensions is an important strategy for generalization.

Quantitative researchers using nonprobability samples must be cautious about the inferences they make. With efforts to deliberately enhance representativeness, a conservative interpretation of the results with regard to generalizability, and replication of the study with new samples, researchers find that nonprobability samples usually work reasonably well.

PROBABILITY SAMPLING

Probability sampling involves the random selection of elements from a population. **Random sampling** involves a selection process in which each element in the population has an equal, independent chance of being selected. Probability sampling is a complex, technical topic, and books such as those by Levy and Lemeshow (2009) offer further guidance for advanced students.

TIP: Random sampling should not be (but often is) confused with random assignment, which was described in connection with experimental designs in Chapter 9. Random assignment is the process of allocating people to different treatment conditions at random. Random *assignment* has no bearing on how people in an RCT were selected in the first place.

Simple Random Sampling

Simple random sampling is the most basic probability sampling design. In simple random sampling, researchers establish a **sampling frame**, the technical name for the list of elements from which the sample will be chosen. If nursing students at the University of Connecticut were the accessible population, then a roster of those students would be the sampling frame. If the sampling unit were 300-bed or larger hospitals in Taiwan, then a list of all such hospitals would be the sampling frame. In practice, a population may be defined in terms of an existing sampling frame. For example, if we wanted to use a voter registration list as a sampling frame, we would have to define the community population as residents who had registered to vote.

Once a sampling frame has been developed, elements are numbered consecutively. A table of random numbers or computer-generated list of random numbers would then be used to draw a sample of the desired size. An example of a sampling frame for a population of 50 people is shown in Table 12.3. Let us assume we want to randomly sample 20 people. As with random assignment, we could find a starting place in a table of random numbers by blindly placing our finger at some point on the page to

TABLE 12.3 Sampling Frame for Simple Random Sampling Example

1.	N. Alexander	**26.**	C. Ball
2.	D. Brady	27.	L. Chodos
3.	D. Carroll	28.	K. DiSanto
4.	M. Dakes	29.	B. Eddy
5.	H. Edelman	**30.**	J. Fishon
6.	L. Forester	**31.**	R. Griffin
7.	J. Galt	32.	B. Hebert
8.	L. Hall	**33.**	C. Joyce
9.	R. Ivry	**34.**	S. Kane
10.	A. Janosy	35.	C. Lace
11.	J. Kettlewell	36.	M. Montanari
12.	L. Lack	37.	B. Nicolet
13.	B. Mastrianni	**38.**	T. Opitz
14.	K. Nolte	39.	J. Portnoy
15.	N. O'Hara	40.	G. Queto
16.	T. Piekarz	41.	A. Ryan
17.	J. Quint	42.	S. Singleton
18.	M. Riggi	**43.**	L. Tower
19.	M. Solomons	44.	V. Vaccaro
20.	S. Thompson	**45.**	B. Wilmot
21.	C. VanWagner	**46.**	D. Abraham
22.	R. Walsh	47.	V. Brusser
23.	J. Yepsen	48.	O. Crampton
24.	M. Zimmerman	49.	R. Davis
25.	A. Arnold	**50.**	C. Eldred

find a two-digit combination between 1 and 50. For this example, suppose that we began with the first number in the random number table of Table 9.2 (p. 208), which is 46. The person corresponding to that number, D. Abraham, is the first person selected to participate in the study. Number 05, H. Edelman, is the second selection, and number 23, J. Yepsen, is the third. This process would continue until 20 participants are chosen. The selected elements are circled in Table 12.3.

Clearly, a sample selected randomly in this fashion is not subject to biases. Although there is no guarantee that a random sample will be representative, random selection ensures that differences in the attributes of the sample and the population are purely a function of chance. The probability of selecting a deviant sample decreases as the size of the sample increases.

Simple random sampling tends to be laborious. Developing a sampling frame, numbering all elements, and selecting elements are time-consuming chores, particularly if the population is large. Imagine enumerating all the telephone subscribers listed in the New York City telephone directory! In actual practice, simple random sampling is not used frequently because it is relatively inefficient. Furthermore, it is not always possible to get a listing of every element in the population, so other methods may be required.

Example of a simple random sample: Lipman and colleagues (2009) documented nurses' practices in an urban children's hospital with regard to whether children's height was measured and plotted on growth charts. Using a random numbers table, a simple random sample of 200 hospital charts was selected for review.

Stratified Random Sampling

In **stratified random sampling**, the population is first divided into two or more strata. As with quota sampling, the aim is to enhance representativeness. Stratified sampling designs subdivide the population into homogeneous subsets (e.g., based on gender or illness severity categories) from which an appropriate number of elements are selected at random.

One difficulty with stratification is that the stratifying attributes must be known in advance and may not be readily discernible. Patient listings, student rosters, or organizational directories may contain information for meaningful stratification, but many lists do not. Quota sampling does not have the same problem because researchers can ask people questions that determine their eligibility for a particular stratum. In stratified sampling, however, a person's status in a stratum must be known before random selection.

The most common procedure for drawing a stratified sample is to group together elements belonging to a stratum and to select randomly the desired number of elements. To illustrate, suppose that the list in Table 12.3 consisted of 25 men (numbers 1 through 25) and 25 women (numbers 26 through 50). Using gender as the stratifying variable, we could guarantee a sample of 10 men and 10 women by randomly sampling 10 numbers from the first half of the list and 10 from the second half. As it turns out, our simple random sampling did result in 10 elements being chosen from each half of the list, but this was purely by chance. It would not have been unusual to draw, say, 8 names from one half and 12 from the other. Stratified sampling can guarantee the appropriate representation of different population segments.

Stratification usually divides the population into unequal subpopulations. For example, if the person's race were used to stratify the population of U.S. citizens, the subpopulation of white people would be larger than that of nonwhite people. We might select participants in proportion to the size of the stratum in the population, using **proportionate stratified sampling**. If the population was students in a nursing school that had 10% African American, 10% Hispanic, 10% Asian, and 70% white students, then a proportionate stratified sample of 100 students, with race/ethnicity as the stratifying variable, would consist of 10, 10, 10, and 70 students from the respective strata.

Proportionate sampling may result in insufficient numbers for making comparisons among strata. In our example, we would not be justified in drawing conclusions about Hispanic nursing students based

on only 10 cases. For this reason, researchers may use **disproportionate sampling** when comparisons are sought between strata of greatly unequal size. In the example, the sampling proportions might be altered to select 20 African American, 20 Hispanic, 20 Asian, and 40 white students. This design would ensure a more adequate representation of the three racial/ethnic minorities. When disproportionate sampling is used, however, it is necessary to make an adjustment to arrive at the best estimate of *overall* population values. This adjustment, called **weighting**, is a simple mathematic computation described in textbooks on sampling.

Stratified random sampling enables researchers to sharpen the representativeness of their samples. When it is desirable to obtain reliable information about subpopulations whose memberships are small, stratification provides a means of including a sufficient number of cases in the sample by oversampling for that stratum. Stratified sampling, however, may be impossible if information on the critical variables is unavailable. Furthermore, a stratified sample requires even more labor and effort than simple random sampling because the sample must be drawn from multiple enumerated listings.

Example of stratified random sampling: Ekwall and Hallberg (2007) studied caregiver satisfaction among informal older caregivers who participated in a mail survey in Sweden. The sample was stratified on the basis of age. Questionnaires were mailed to 2,500 elders aged 75 to 79, 2,500 elders aged 80 to 84, 2,000 elders aged 85 to 89, and 1,500 elders aged 90 and over.

Multistage Cluster Sampling

For many populations, it is impossible to get a listing of all elements. For example, the population of full-time nursing students in the United Kingdom would be difficult to list and enumerate for the purpose of drawing a simple or stratified random sample. Large-scale surveys—especially ones involving personal interviews—almost never use simple or stratified random sampling; they usually rely on multistage sampling, beginning with clusters.

Cluster sampling involves selecting broad groups (clusters) rather than selecting individuals, and is typically the first stage of a multistage approach. In drawing a sample of nursing students, we might first draw a random sample of nursing schools and then draw a sample of students from the selected schools. The usual procedure for selecting samples from a general population in the United States is to sample successively such administrative units as census tracts, then households, and then household members. The resulting design can be described in terms of the number of stages (e.g., three-stage sampling). Clusters can be selected either by simple or stratified methods. For instance, in selecting clusters of nursing schools, it may be advisable to stratify on program type.

For a specified number of cases, multistage sampling tends to be less accurate than simple or stratified random sampling. Yet, multistage sampling is more practical than other types of probability sampling, particularly when the population is large and widely dispersed.

Example of multistage sampling: Callaghan and colleagues (2010) studied self-efficacy and exercise behavior in a large sample of Chinese students. High schools were first sampled, with stratification based on geographic location. Students were subsequently sampled from the selected high schools.

Systematic Sampling

Systematic sampling involves selecting every *k*th case from a list, such as every 10th person on a patient list or every 25th person on a student roster. Systematic sampling is sometimes used to sample every *k*th person entering a store, or passing down the street, or leaving a hospital, and so forth. In such situations, unless the population is narrowly defined as all those people entering, passing by, or leaving, the sampling is essentially a sample of convenience.

Systematic sampling can, however, be applied so that an essentially random sample is drawn. If we had a list (sampling frame), the following procedure could be adopted. The desired sample size

is established at some number (n). The size of the population must be known or estimated (N). By dividing N by n, the sampling interval width (k) is established. The **sampling interval** is the standard distance between sampled elements. For instance, if we wanted a sample of 200 from a population of 40,000, then our sampling interval would be as follows:

$$k = \frac{40{,}000}{200} = 200$$

In other words, every 200th element on the list would be sampled. The first element should be selected randomly. Suppose that we randomly selected number 73 from a random number table. People corresponding to numbers 73, 273, 473, and so on would be sampled. Alternatively, we could randomly select a number from 1 to the number of elements listed on a page, and then randomly select every *k*th unit on all pages (e.g., number 38 on every page).

Systematic sampling conducted in this manner yields essentially the same results as simple random sampling, but involves less work. Problems would arise if the list were arranged in such a way that a certain type of element is listed at intervals coinciding with the sampling interval. For instance, if every 10th nurse listed in a nursing staff roster was a head nurse and the sampling interval was 10, then head nurses would either always or never be included in the sample. Problems of this type are rare, fortunately. Systematic sampling may be preferred to simple random sampling because similar results are obtained in a more efficient manner. Systematic sampling can also be applied to lists that have been stratified.

Example of a systematic sample: Houghton and colleagues (2008) surveyed nurse anesthetists about their practices and attitudes regarding smoking intervention. Using the membership list of the American Association of Nurse Anesthetists, every 30th name in the alphabetized list was selected for the sample.

Evaluation of Probability Sampling

Probability sampling is the best method of obtaining representative samples. If all the elements in a population have an equal probability of being selected, then the resulting sample is likely to do a good job of representing the population. A further advantage is that probability sampling allows researchers to estimate the magnitude of sampling error. **Sampling error** refers to differences between population values (such as the average age of the population) and sample values (such as the average age of the sample).

The great drawback of probability sampling is its impracticality. It is beyond the scope of most studies to involve a probability sample, unless the population is narrowly defined—and if it *is* narrowly defined, probability sampling may be "overkill." Probability sampling is the preferred and most respected method of obtaining sample elements, but is often unfeasible.

TIP: The quality of the sampling plan is of particular importance in survey research, because the purpose of surveys is to obtain information about the prevalence or average values for a population. All national surveys, such as the National Health Interview Survey in the United States, use probability samples (usually multistage cluster samples). Probability samples are rarely used in experimental and quasi-experimental studies, in part because the main focus of such inquiries is on between-group differences rather than absolute values for a population.

SAMPLE SIZE IN QUANTITATIVE STUDIES

Quantitative researchers need to pay attention to the number of participants needed to achieve statistical conclusion validity. A procedure called **power analysis** (Cohen, 1988) can be used to estimate sample size needs, but some statistical knowledge is needed before this procedure can be explained. In this section, we offer guidelines to beginning researchers; advanced students can read about power analysis in Chapter 17 or in a sampling or statistics textbook (e.g., Polit, 2010).

Sample Size Basics

There are no simple formulas that can tell you how large a sample you will need in a given study, but as

a general recommendation, you should use the largest sample possible. The larger the sample, the more representative of the population it is likely to be. Every time researchers calculate a percentage or an average based on sample data, they are estimating a population value. Smaller samples tend to produce less precise estimates than larger ones. In other words, the larger the sample, the smaller the sampling error.

Let us illustrate this with an example of monthly aspirin consumption in a nursing home (Table 12.4). The population consists of 15 residents whose aspirin consumption averages 16.0 aspirins per month, as shown in the top row of the table. Eight simple random samples—two each with sample sizes of 2, 3, 5, and 10—have been drawn. Each sample average represents an estimate of the population average (i.e., 16.0). With a sample size of two, our estimate might have been wrong by as many as eight aspirins (sample 1B, average of 24.0), which is 50% greater than the population value. As the sample size increases, the averages get closer to the true population value, *and* the differences in the estimates between samples A and B get smaller as well. As sample size increases, the probability of getting a markedly deviant sample diminishes. Large samples provide an opportunity to counterbalance atypical values. In the absence of a power analysis, the safest procedure is to obtain data from as large a sample as is feasible.

Large samples are no assurance of accuracy, however. When nonprobability sampling methods are used, even a large sample can harbor extensive bias. The famous example illustrating this point is the 1936 American presidential poll conducted by the magazine *Literary Digest*, which predicted that Alfred M. Landon would defeat Franklin D. Roosevelt by a landslide. About 2.5 million individuals participated in this poll—a substantial sample. Biases resulted from the fact that the sample was drawn from telephone directories and automobile registrations during a depression year when only the well-to-do (who preferred Landon) had a car or telephone. Thus, a large sample cannot correct for a faulty sampling design. Nevertheless, a large nonprobability sample is preferable to a small one.

Because practical constraints such as time, participant cooperation, and resources often limit sample

TABLE 12.4 Comparison of Population and Sample Values and Averages: Nursing Home Aspirin Consumption Example

NUMBER OF PEOPLE IN GROUP	GROUP	INDIVIDUAL DATA VALUES (NUMBER OF ASPIRINS CONSUMED, PRIOR MONTH)	AVERAGE
15	Population	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30	16.0
2	Sample 1A	6, 14	10.0
2	Sample 1B	20, 28	24.0
3	Sample 2A	16, 18, 8	14.0
3	Sample 2B	20, 14, 26	20.0
5	Sample 3A	26, 14, 18, 2, 28	17.6
5	Sample 3B	30, 2, 26, 10, 4	14.4
10	Sample 4A	22, 16, 24, 20, 2, 8, 14, 28, 20, 4	15.8
10	Sample 4B	12, 18, 8, 10, 16, 6, 28, 14, 30, 22	16.4

size, many nursing studies are based on relatively small samples. Most nursing studies use samples of convenience, and many are based on samples that are too small to provide an adequate test of the research hypotheses. Quantitative studies usually are based on samples of fewer than 200 participants, and many have fewer than 100 people (e.g., Polit & Sherman, 1990; Polit & Gillespie, 2009). Power analysis is not done routinely by nurse researchers, and research reports often offer no justification for sample size. When samples are too small, quantitative researchers run the risk of gathering data that will not support their hypotheses, *even when their hypotheses are correct*, thereby undermining statistical conclusion validity.

Factors Affecting Sample Size Requirements in Quantitative Research

Sample size requirements are affected by various factors, some of which we discuss in this section.

Effect Size

Power analysis builds on the concept of an **effect size**, which expresses the strength of relationships among research variables. If there is reason to expect that the independent and dependent variables will be strongly related, then a relatively small sample may be adequate to reveal the relationship statistically. For example, if we were testing a powerful new drug to treat AIDS, it might be possible to demonstrate its effectiveness with a small sample. Typically, however, nursing interventions have modest effects, and variables are usually only moderately correlated with one another. When there is no *a priori* reason for believing that relationships will be strong, then small samples are risky.

Homogeneity of the Population

If the population is relatively homogeneous, a small sample may be adequate. The greater the variability, the greater is the risk that a small sample will not adequately capture the full range of variation. For most nursing studies, it is probably best to assume a fair degree of heterogeneity, unless there is evidence from prior research to the contrary.

Cooperation and Attrition

In most studies, not every one invited to participate in a study agrees to do so. Therefore, in developing a sampling plan, it is good to begin with a realistic, evidence-based estimate of the percentage of people likely to cooperate. Thus, if your targeted sample size is 200 but you expect a 50% refusal rate, you would have to recruit 400 or so eligible people.

In studies with multiple points of data collection, the number of participants usually declines over time. Attrition is most likely to occur if the time lag between data collection points is great, if the population is mobile, or if the population is at risk of death or disability. If the researcher has an ongoing relationship with participants (as might be true in clinical studies), then attrition might be low—but it is rarely 0%. Therefore, in estimating sample size needs, researchers should factor in anticipated loss of participants over time.

Attrition problems are not restricted to longitudinal studies. People who initially agree to cooperate in a study may be subsequently unable or unwilling to participate for various reasons, such as death, deteriorating health, early discharge, discontinued need for an intervention, or simply a change of heart. Researchers should expect a certain amount of participant loss and recruit accordingly.

TIP: Polit and Gillespie (2009) found, in a sample of over 100 nursing RCTs, that the average participant loss was 12.5% for studies with follow-up data collection between 31 and 90 days after baseline, and was 18% when the final data collection was more than 6 months after baseline.

Subgroup Analyses

Researchers sometimes wish to test hypotheses not only for an entire population, but also for subgroups. For example, we might be interested in assessing whether a structured exercise program is effective in improving infants' motor skills. After testing the general hypothesis with a sample of infants, we might wish to test whether the intervention is more effective for certain infants (e.g., low-birth-weight versus normal-birth-weight infants). When a sample is divided to test for **subgroup effects**, the sample

must be large enough to support analyses with such divisions of the sample.

Sensitivity of the Measures

Instruments vary in their ability to measure key concepts precisely. Biophysiologic measures are usually very sensitive—they measure phenomena accurately, and can make fine discriminations in values. Psychosocial measures often contain some error and lack precision. When measuring tools are imprecise and susceptible to errors, larger samples are needed to test hypotheses adequately.

TIP: Hertzog (2008) has offered guidance on estimating sample size needs for pilot studies.

IMPLEMENTING A SAMPLING PLAN IN QUANTITATIVE STUDIES

This section provides some practical guidance about implementing a sampling plan.

Steps in Sampling in Quantitative Studies

The steps to be undertaken in drawing a sample vary somewhat from one sampling design to the next, but a general outline of procedures can be described.

1. *Identify the population*. You should begin with a clear idea about the target population to which you would like to generalize your results. Unless you have extensive resources, you are unlikely to have access to the entire target population, so you will also need to identify the population that is accessible to you. Researchers sometimes *begin* by identifying an accessible population, and then decide how best to characterize the target population.
2. *Specify the eligibility criteria*. The criteria for eligibility in the sample should then be spelled out. The criteria should be as specific as possible with regard to characteristics that might exclude potential participants (e.g., extremes of poor health, inability to read English). The criteria might lead you to redefine your target population.
3. *Specify the sampling plan*. Once the accessible population has been identified, you must decide (a) the method of drawing the sample and (b) how large it will be. Sample size specifications should consider the aspects of the study discussed in the previous section. If you can perform a power analysis to estimate the needed number of participants, we highly recommend that you do so. Similarly, if probability sampling is a viable option, that option should be exercised. If you are not in a position to do either, we recommend using as large a sample as possible and taking steps to build representativeness into the design (e.g., by using quota or consecutive sampling).
4. *Recruit the sample*. Once the sampling design has been specified, the next step is to recruit prospective participants according to the plan (after any needed institutional permissions have been obtained) and ask for their cooperation. Issues relating to participant recruitment are discussed next.

Sample Recruitment

Recruiting people to participate in a study involves two major tasks: identifying eligible candidates and persuading them to participate. Researchers may need to spend time early in the project deciding the best sources for recruiting potential participants. Researchers must ask such questions as, Where do large numbers of people matching my population construct live or obtain care? Will I have direct access to people, or will I need to work through gatekeepers? Will there be sufficiently large numbers in one location, or will multiple sites be necessary? During the recruitment phase, it may be necessary to develop a **screening instrument**, which is a brief interview or form that allows researchers to determine whether a prospective participant meets all eligibility criteria for the study.

The next task involves gaining the cooperation of people who have been deemed eligible. It is critical to have an effective recruitment strategy. Many people, given the right circumstances, will agree to cooperate, but—especially in intervention research—some are hesitant. Researchers should ask themselves, What will make this research experience enjoyable, worthwhile, convenient, pleasant, and nonthreatening for people? Researchers have control over such influential factors as the following:

- *Recruitment method*. Face-to-face recruitment is usually more effective than solicitation by a telephone call, letter, or email.
- *Courtesy*. Successful recruitment depends on using recruiters who are pleasant, courteous, and enthusiastic about the study. Cooperation sometimes is enhanced if the recruiters' characteristics are similar to those of prospective participants—particularly with regard to gender, race, and ethnicity.
- *Persistence*. Although high-pressure tactics are never acceptable, persistence may sometimes be needed. When prospective participants are first approached, their initial reaction may be to decline if they are taken off guard. If a person hesitates or gives an equivocal answer at the first attempt, recruiters should ask if they could come back at a later time.
- *Incentives*. Gifts and monetary incentives have been found to have a substantial effect on participation (Edwards et al., 2009).
- *Benefits*. The benefits of participating to the individual and to society should be explained, without exaggeration or misleading information.
- *Sharing results*. Sometimes it is useful to provide people with tangible evidence of their contribution to the study by offering to send them a brief summary of the study results.
- *Convenience*. Every effort should be made to collect data at a time and location that is convenient for participants. In some cases, this may mean making arrangements for transportation or for the care of young children.
- *Endorsements*. It may be valuable to have the study endorsed by a person or organization that has prospective participants' confidence, and to communicate this to them. Endorsements might come from the institution serving as the research setting, a funding agency, or a respected community group or person, such as a church leader. A statement of university sponsorship has positive effects of participation (Edwards et al., 2009). Press releases in advance of recruitment may be advantageous.
- *Assurances*. Prospective subjects should be told who will see the data, what use will be made of the data, and how confidentiality will be maintained.

The issue of participant recruitment—and retention—has received considerable attention in recent years. There are numerous articles on strategies for, and barriers to, recruiting from minority or vulnerable populations (e.g., Russell et al., 2008; Topp et al., 2008; UyBico et al., 2007; Webb et al., 2009), which is a particularly important issue for those interested in health disparities research. Guidance also is available with regard to participant recruitment for RCTs (e.g., Berger et al., 2007; Gul & Ali, 2010; Leathem et al., 2009). In the United States, researchers should be aware of potential recruitment difficulties that have arisen within the context of the Health Insurance Portability and Accountability Act or HIPAA (Wipke-Tevis & Pickett, 2008).

TIP: Participant recruitment often proceeds at a slower pace than researchers anticipate. Once you have determined your sample size needs, it is useful to develop contingency plans for recruiting more people, should the initial plan prove overly optimistic. For example, a contingency plan might involve relaxing the eligibility criteria, identifying another institution through which participants could be recruited, offering incentives to make participation more attractive, or lengthening the recruitment period. When such plans are developed at the outset, it reduces the likelihood that you will have to settle for a less-than-desirable sample size.

Generalizing From Samples

Ideally, the sample is representative of the accessible population, and the accessible population is representative of the target population. By using an appropriate sampling plan, researchers can be reasonably confident that the first part of this ideal has been realized. The second part of the ideal entails greater risk. Are diabetic patients in Atlanta representative of diabetic patients in the United States? Researchers must exercise judgment in assessing the degree of similarity.

The best advice is to be realistic and conservative, and to ask challenging questions: Is it reasonable to assume that the accessible population is representative of the target population? In what ways might they differ? How would such differences affect the conclusions? If differences are great, it would be prudent to specify a more restricted target population to which the findings could be meaningfully generalized.

Interpretations about the generalizability of findings can be enhanced by comparing sample characteristics with population characteristics, when this is possible. Published information about the characteristics of many populations may be available to help in evaluating sampling bias. For example, if you were studying low-income children in Chicago, you could obtain information on the Internet about salient characteristics (e.g., race/ethnicity, age distribution) of low-income American children from the U.S. Bureau of the Census. Population characteristics could then be compared with sample characteristics, and differences taken into account in interpreting the findings. Sousa and colleagues (2004) provide suggestions for drawing conclusions about whether a convenience sample is representative of the population.

Example of comparison of characteristics: Griffin and colleagues (2008) conducted a survey of over 300 pediatric nurses, whose names had been randomly sampled from a list of 9,000 nurses who subscribed to pediatric nursing journals. Demographic characteristics of the sample (e.g., gender, race/ethnicity, educational background) were compared with characteristics of a nationally representative sample of nurses who participated in a government survey.

CRITIQUING SAMPLING PLANS

In coming to conclusions about the quality of evidence that a study yields, you should carefully scrutinize the sampling plan. If the sample is seriously biased or too small, the findings may be misleading or just plain wrong.

You should consider two issues in your critique of a study's sampling plan. The first is whether the researcher adequately described the sampling strategy. Ideally, research reports should include a description of the following:

- The type of sampling approach used (e.g., convenience, simple random)
- The study population and eligibility criteria for sample selection
- The number of participants and a rationale for the sample size, including whether a power analysis was performed
- A description of the main characteristics of sample members (e.g., age, gender, medical condition, and so forth) and, ideally, of the population
- The number and characteristics of potential participants who declined to participate in the study

If the description of the sample is inadequate, you may not be in a position to deal with the second and principal issue, which is whether the researcher made good sampling decisions. And, if the description is incomplete, it will be difficult to draw conclusions about whether the evidence can be applied in your clinical practice.

Sampling plans should be scrutinized with respect to their effects on the construct, internal, external, and statistical conclusion validity of the study. If a sample is small, statistical conclusion validity will likely be undermined. If the eligibility criteria are restrictive, this could benefit internal validity—but possibly to the detriment of construct and external validity.

We have stressed that a key criterion for assessing the adequacy of a sampling plan in quantitative research is whether the sample is representative of the population. You will never know for sure, but if the sampling strategy is weak or if the sample size

BOX 12.1 Guidelines for Critiquing Quantitative Sampling Designs

1. Is the study population identified and described? Are eligibility criteria specified? Are the sample selection procedures clearly delineated?
2. Do the sample and population specifications support an inference of construct validity with regard to the population construct?
3. What type of sampling plan was used? Would an alternative sampling plan have been preferable? Was the sampling plan one that could be expected to yield a representative sample?
4. If sampling was stratified, was a useful stratification variable selected? If a consecutive sample was used, was the time period long enough to address seasonal or temporal variation?
5. How were people recruited into the sample? Does the method suggest potential biases?
6. Did some factor other than the sampling plan (e.g., a low response rate) affect the representativeness of the sample?
7. Are possible sample biases or weaknesses identified by the researchers themselves?
8. Are key characteristics of the sample described (e.g., mean age, percent female)?
9. Is the sample size sufficiently large to support statistical conclusion validity? Was the sample size justified on the basis of a power analysis or other rationale?
10. Does the sample support inferences about external validity? To whom can the study results reasonably be generalized?

is small, there is reason to suspect some bias. When researchers adopt a sampling plan in which the risk for bias is high, they should take steps to estimate the direction and degree of this bias so that readers can draw some informed conclusions.

Even with a rigorous sampling plan, the sample may be biased if not all people invited to participate in a study agree to do so—which is almost always the case. If certain segments of the population refuse to participate, then a biased sample can result, even when probability sampling is used. Research reports ideally should provide information about **response rates** (i.e., the number of people participating in a study relative to the number of people sampled), and about possible **nonresponse bias**—differences between participants and those who declined to participate (also sometimes referred to as *response bias*). In longitudinal studies, attrition bias should be reported.

Quantitative researchers make decisions about the specification of the population as well as the selection of the sample. If the target population is defined broadly, researchers may have missed opportunities to control confounding variables, and the gap between the accessible and the target population may be too great. One of your jobs as reviewer is to come to conclusions about the reasonableness of generalizing the findings from the researcher's sample to the accessible population and from the accessible population to a broader target population. If the sampling plan is seriously flawed, it may be risky to generalize the findings at all without replicating the study with another sample.

Box 12.1 presents some guiding questions for critiquing the sampling plan of a quantitative research report.

RESEARCH EXAMPLE

In this section, we describe in some detail the sampling plan of a quantitative nursing study.

Studies: (1) Quality and strength of patient safety climate on medical–surgical units (Hughes et al., 2009); (2) Organizational effects on patient satisfaction in hospital medical–surgical units (Bacon & Mark, 2009); and (3) Nurse staffing and medication errors: Cross-sectional or longitudinal relationships? (Mark & Belyea, 2009).

Purpose: Barbara Mark, with funding from NINR, launched a large multisite study called the Outcomes Research in Nursing Administration Project-II (ORNA-II). The overall purpose was to investigate relationships of hospital context and structure on the one hand and patient, nurse, and organization outcomes on the other. Data from this project have been used in numerous studies, three of which are cited here.

Design: The project was designed as a prospective correlational study, with data collected in 2003 and 2004.

Sampling Plan: Sampling was multistaged. In the first stage, 146 acute care hospitals were randomly selected from a list of hospitals accredited by the Joint Commission on Accreditation of Health Organizations. To be included, hospitals had to have at least 99 licensed beds. Hospitals were excluded if they were federal, for-profit, or psychiatric facilities. Then, from each selected hospital, two medical, surgical, or medical–surgical units were selected to participate in the study. Units were excluded if they were critical care, pediatric, obstetric, or psychiatric units. Among hospitals with only two eligible units, both participated. Among hospitals with more than two eligible units, an on-site study coordinator selected two to participate. Ultimately, 281 nursing units in 143 hospitals participated in the study. Data from each hospital were gathered in three rounds of data collection over a 6-month period. On each participating unit, all RNs with more than 3 months of experience on that unit were asked to respond to three sets of questionnaires. The response rates were 75% of nurses at Time 1 (4,911 nurses), 58% at Time 2 (3,689 nurses), and 53% at Time 3 (3,272 nurses). Patients were also invited to participate at Time 3. Ten patients on each unit were randomly selected to complete a questionnaire. Patients were included if they were 18 years of age or older, had been hospitalized for at least 48 hours, were able to speak and read English, and were not scheduled for immediate discharge. A total of 2,720 patients participated, and the response rate was 91%.

Key Findings:

- Nurses in Magnet hospitals were more likely to communicate about errors and participate in error-related problem solving (Hughes et al., 2009)
- Greater availability of nursing unit support services was associated with higher levels of patient satisfaction (Bacon & Mark, 2009)
- Nurse staffing was unrelated to medication errors (Mark & Belyea, 2009)

SUMMARY POINTS

- **Sampling** is the process of selecting a portion of the **population**, which is an entire aggregate of cases. An **element** is the basic population unit about which information is collected—usually humans in nursing research.
- **Eligibility criteria** are used to establish population characteristics and to determine who could participate in a study—either who can be included (**inclusion criteria**) or who should be excluded (**exclusion criteria**). Care must be taken to specify eligibility criteria so as to maximize the construct validity of the population construct.
- Researchers usually sample from an **accessible population**, but should identify the **target population** to which they want to generalize their results.
- A sample in a quantitative study is assessed in terms of **representativeness**—the extent to which the sample is similar to the population and avoids bias. **Sampling bias** refers to the systematic over-representation or under-representation of some segment of the population.
- Methods of **nonprobability sampling** (wherein elements are selected by nonrandom methods) include convenience, quota, consecutive, and purposive sampling. Nonprobability sampling designs are practical but usually have strong potential for bias.
- **Convenience sampling** uses the most readily available or convenient group of people for the sample. **Snowball sampling** is a type of convenience sampling in which referrals for potential participants are made by those already in the sample.
- **Quota sampling** divides the population into homogeneous **strata** (subpopulations) to ensure representation of subgroups; within each stratum, people are sampled by convenience.
- **Consecutive sampling** involves taking *all* of the people from an accessible population who meet the eligibility criteria over a specific time interval, or for a specified sample size.

- In **purposive sampling**, elements are handpicked to be included in the sample based on the researcher's knowledge about the population.
- **Probability sampling** designs, which involve the random selection of elements from the population, yield more representative samples than nonprobability designs and permit estimates of the magnitude of **sampling error**.
- **Simple random sampling** involves the random selection of elements from a **sampling frame** that enumerates all the elements; **stratified random sampling** divides the population into homogeneous strata from which elements are selected at random.
- **Cluster sampling** involves sampling of large units. In **multistage sampling**, there is a successive, multistaged selection of random samples from larger units (clusters) to smaller units (individuals) by either simple random or stratified random methods.
- **Systematic sampling** is the selection of every *k*th case from a list. By dividing the population size by the desired sample size, the researcher establishes the **sampling interval**, which is the standard distance between the selected elements.
- In quantitative studies, researchers should use a **power analysis** to estimate **sample size** needs. Large samples are preferable to small ones because larger samples enhance statistical conclusion validity and tend to be more representative, but even large samples do not *guarantee* representativeness.

STUDY ACTIVITIES

Chapter 12 of the *Resource Manual for Nursing Research*: *Generating and Assessing Evidence for Nursing Practice, 9th edition*, offers exercises and study suggestions for reinforcing concepts presented in this chapter. In addition, the following study questions can be addressed:

1. Answer relevant questions from Box 12.1 with regard to sampling plan for the ORNA studies, described at the end of the chapter. Also consider the following additional questions: (a) How many stages would you say were involved in the sampling plan? (b) What are some of the likely sources of sampling bias in the final sample of 3,272 nurses?
2. Use the table of random numbers in Table 9.2 to select 10 names from the list of people in Table 12.3. How many names did you draw from the first 25 names and from the second 25 names?

STUDIES CITED IN CHAPTER 12

Bacon, C. T., & Mark, B. (2009). Organizational effects on patient satisfaction in hospital medical-surgical units. *Journal of Nursing Administration, 39*, 220–227.

Callaghan, P., Khalil, E., & Morres, I. (2010). A prospective evaluation of the Transtheoretical Model of Change applied to exercise in young people. *International Journal of Nursing Studies, 47*, 3–12.

Dudley-Brown, S., & Freivogel, M. (2009). Hereditary colorectal cancer in the gastroenterology clinic: How common are at-risk patients and how do we find them? *Gastroenterology Nursing, 32*, 8–16.

Ekwall, A., & Hallberg, I. (2007). The association between caregiving satisfaction, difficulties and coping among older family caregivers. *Journal of Clinical Nursing, 16*, 832–844.

Fox, M., Sidani, S., & Brooks, D. (2009). Perceptions of bed days for individuals with chronic illness in extended care facilities. *Research in Nursing & Health, 32*, 335–344.

Griffin, R., Polit, D., & Byrne, M. (2008). Nurse characteristics and inferences about children's pain. *Pediatric Nursing, 34*, 297–305.

Hafsteindóttir, T., Mosselman, M., Schoneveld, C., Riedstra, Y., & Kruitwagen, C. (2010). Malnutrition is hospitalized neurological patients approximately doubles in 10 days of hospitalisation. *Journal of Clinical Nursing, 19*, 639–645.

Houghton, C., Marcukaitis, A., Marienau, M., Hooten, M., Stevens, S., & Warner, D. (2008). Tobacco intervention attitudes and practices among certified registered nurse anesthetists. *Nursing Research, 57*, 123–129.

Hughes, L., Chang, Y., & Mark, B. (2009). Quality and strength of patient safety climate on medical-surgical units. *Health Care Management Review, 34*, 19–28.

Lipman, T., Euler, D., Markowitz, G., & Ratclifee, S. (2009). Evaluation of linear measurement and growth plotting in an

inpatient pediatric setting. *Journal of Pediatric Nursing, 24*, 323–329.

Mark, B., & Belyea, M. (2009). Nurse staffing and medication errors: Cross-sectional or longitudinal relationships? *Research in Nursing & Health, 32*, 18–30.

O'Meara, D., Mireles-Cabodevila, E., Frame, E., Hummell, A., Hammel, J., Dweik, R., & Arroliga, A. C. (2008). Evaluation of delivery of enteral nutrition in critically ill patients receiving mechanical ventilation. *American Journal of Critical Care, 17*, 53–61.

Peddle, C., Jones, L., Eves, N., Reiman, T. Sellar, C., Winton, T., & Courneya, K. S. (2009). Correlates of adherence to supervised exercise in patients awaiting surgical removal of malignant lung lesions. *Oncology Nursing Forum, 36*, 287–295.

Methodologic and nonresearch references cited in this chapter can be found in a separate section at the end of the book.

13 Data Collection in Quantitative Research

Quantitative researchers collect data in a structured manner. Both the people collecting the data and the study participants are constrained during the collection of structured data. Constraints are imposed so that there is consistency in what is asked and how answers are reported, in an effort to enhance objectivity, reduce biases, and facilitate analysis. Major methods of collecting structured data are discussed in this chapter. We begin by discussing broad planning issues.

DEVELOPING A DATA COLLECTION PLAN

Data collection plans for quantitative studies ideally yield accurate, valid, and meaningful data. This is a challenging goal, typically requiring considerable time and effort to achieve. Steps in developing a data collection plan are described in this section. (A flowchart illustrating the sequence of steps is available in the Toolkit of the accompanying *Resource Manual*.)

Identifying Data Needs

Researchers usually begin by identifying the types of data needed for their study. Advance planning may help to avoid "if only" disappointments at the analysis stage. In quantitative studies, researchers may need data for the following purposes:

1. *Testing hypotheses or addressing research questions*. Researchers must include one or more measures of all key variables. Multiple measures of some variables may be needed if a variable is complex or if there is an interest in corroboration and triangulation.
2. *Describing sample characteristics*. Information should be gathered about major demographic and health characteristics of the sample. We advise gathering data about participants' age, gender, race or ethnicity, and education (or income). This information is critical in interpreting results and understanding the population to whom findings can be generalized. If the sample includes participants with a health problem, data on the nature of that problem also should be gathered (e.g., severity, treatments, time since diagnosis).

TIP: Asking demographic questions in the right way is more difficult than you might think. Because the need to collect information about sample characteristics is nearly universal, we have included a demographic form and guidelines in the Toolkit of the accompanying *Resource Manual*. The demographic questionnaire can be adapted as needed.

3. *Controlling confounding variables*. Various approaches can be used to control confounding variables, many of which require measuring those variables. For example, for analysis of covariance, variables that are statistically controlled must be measured.
4. *Analyzing potential biases*. Data that can help the researcher to identify potential biases should be collected. For example, researchers should gather information that would help to identify selection biases in a nonequivalent control group design or attrition biases in RCTs.
5. *Understanding subgroup effects*. It is often desirable to answer research questions for key subgroups of participants. For example, we may wish to know if a special intervention for indigent pregnant women is equally effective for primiparas and multiparas. In such a situation, we would need to collect data about the participants' childbearing history.
6. *Interpreting results*. Researchers should try to anticipate alternative results, and then assess what types of data would best help in interpreting them. For example, if we hypothesized that the presence of school-based clinics in high schools would lower the incidence of sexually transmitted diseases among students but found that the incidence remained constant after the clinic opened, what type of information would help us interpret this result (e.g., information about the students' frequency of intercourse, number of partners, use of condoms, and so on)?
7. *Assessing treatment fidelity*. In intervention studies, it is often useful to monitor treatment fidelity and to assess whether the intended treatment was actually received.
8. *Obtaining administrative information*. It is usually necessary to gather administrative data—for example, dates of data collection and contact information in longitudinal studies.

The list of possible data needs may seem daunting, but many categories overlap. For example, participant characteristics for sample description are often key confounding variables, or useful in creating subgroups. If time or resource constraints make it impossible to collect the full range of variables, then researchers should prioritize data needs.

TIP: In prioritizing data needs, it may be useful to develop a matrix so that decisions about data collection strategies can be made in a systematic way. Such a matrix can help to identify "holes" and redundancies. The matrix might contain such column headings as variable name, purpose (e.g., from the above list), name of instrument to be used, and data quality. A partial example of such a matrix is included in the Toolkit of the *Resource Manual* for you to use and adapt. A conceptual map (Chapter 6) is also a useful tool in identifying data needs.

Selecting Types of Measures

After data needs have been identified, the next step is to select a data collection method (e.g., self-report, records) for each variable. In reviewing data needs, researchers should determine how best to capture each variable in terms of its conceptual or theoretical definition. It is not unusual to combine self-reports, observations, physiologic, or records data in a single study.

Research needs are not the only factors that drive decisions about data collection methods. The decisions must also be guided by ethical considerations (e.g., whether covert data collection is warranted), cost constraints, availability of assistants to help with data collection, and other issues discussed in the next section. Data collection is often the costliest and most time-consuming portion of a study. Because of this, researchers often have to make a number of compromises about the type or amount of data collected.

Selecting and Developing Instruments

Once preliminary decisions have been made about the data collection methods, researchers should determine if there are instruments available for measuring study variables, as will often be the case. Potential data collection instruments should then be assessed. The primary consideration is conceptual relevance: Does the instrument correspond

to your conceptual definition of the variable? Another important criterion is whether the instrument will yield high-quality data. Approaches to evaluating data quality are discussed in Chapter 14. Additional factors that may affect your decisions in selecting an instrument are as follows:

1. *Resources*. Resource constraints sometimes prevent the use of the highest-quality measures. There may be some direct costs associated with the measure (e.g., some psychological tests must be purchased), but the biggest cost involves compensation to data collectors if you cannot do it single-handedly—that is, if you have to hire interviewers or observers. In such a situation, the instrument's administration time may determine whether it is a viable option. Also, it may be necessary to pay a participant stipend if data collection procedures are burdensome. Data collection costs should be carefully considered, especially if the use of expensive methods means that you will be forced to cut costs elsewhere (e.g., using a smaller sample).
2. *Availability and familiarity*. You may need to consider how readily available or accessible various instruments are, especially biophysiologic ones. Similarly, data collection strategies with which you have had experience are usually preferable to new ones because administration is usually smoother and more efficient in such cases.
3. *Population appropriateness*. Instruments must be chosen with the characteristics of the target population in mind. Characteristics of special importance include participants' age and literacy levels. If there is concern about participants' reading skills, it may be necessary to calculate the readability of a prospective instrument. If participants include members of minority groups, you should strive to find instruments that are culturally appropriate. If non–English-speaking participants are included in the sample, then the selection of an instrument may be based on the availability of a translated version.
4. *Norms and comparisons*. It may be desirable to select an instrument that has relevant norms. **Norms** indicate the "normal" values on the measure for a specified population, and thus offer a built-in comparison. Many standardized scales (e.g., the SF-36 Health Survey from the Medical Outcomes Study) have norms. Similarly, it may be advantageous to select an instrument because it was used in other similar studies, thus providing useful information for interpreting study findings. When a study is an intentional replication, it is often important to use the same instruments as in the original study, even if higher-quality measures are available.
5. *Administration issues*. Some instruments have special requirements that need to be considered. For example, obtaining information about the developmental status of children sometimes requires the skills of a professional psychologist. Another administration issue is that some instruments require or assume stringent conditions with regard to the time of administration, privacy of the setting, and so on. In such a case, requirements for obtaining valid measures must match attributes of the research setting.
6. *Reputation*. Instruments designed to measure the same construct often differ in the reputation they enjoy among specialists in a field, even if they are comparable with regard to documented quality. Thus, it may be useful to seek the advice of knowledgeable people, preferably ones with personal, direct experience using the instruments.

If existing instruments are not suitable for some variables, you may be faced with either adapting an instrument or developing a new one. Creating a new instrument should be a last resort, especially for novice researchers, because it is challenging to develop accurate and valid measuring tools. Chapter 15 provides guidance on developing self-report instruments.

If you are fortunate in identifying a suitable instrument, your next step likely will be to obtain written permission from the author to use it. In general, copyrighted materials always require

permission. Instruments that have been developed under a government grant are usually in the public domain, and so may not require permission. When in doubt, it is best to obtain permission. By contacting the instrument's author for permission, you can also request more information about the instrument and its quality. (A sample letter requesting permission to use an instrument is in the Toolkit.)

TIP: In finalizing decisions about instruments, it may be necessary to balance trade-offs between data quality and data quantity (i.e., the number of instruments or questions). If compromises have to be made, it is usually preferable to forego quantity.

Pretesting the Data Collection Package

Researchers who develop a new instrument usually subject it to rigorous **pretesting** so that it can be evaluated and refined. Even when the data collection plan involves existing instruments, however, it is wise to conduct a small pretest.

One purpose of a pretest is to see how much time it takes to administer the entire instrument package. Typically, researchers use multiple instruments and it may be difficult to estimate how long it will take to administer the complete set. Time estimates may be required for informed consent purposes, for developing a budget, or for assessing participant burden.

Pretests can serve many other purposes, including the following:

- Identifying parts of the instrument package that are difficult for participants to read or understand or that may have been misinterpreted
- Identifying questions that participants find objectionable or offensive
- Assessing whether the sequencing of questions or instruments is sensible
- Evaluating training needs for data collectors
- Determining if the measures yield data with sufficient variability

The last purpose requires explanation. For most research questions, the instruments ideally discriminate among participants with different levels of an attribute. If we are asking, for example, whether women experience greater depression than men when they learn of a cancer diagnosis, we need an instrument capable of distinguishing between people with higher and lower levels of depression. If an instrument yields data with limited variability, then it will be impossible to detect a difference in depression between men and women—even when such a difference actually exists. Thus, researchers should look at pretest variation on key research variables. To pursue the example, if the entire pretest sample looks very depressed (or not at all depressed), it would probably be necessary to pretest another instrument.

Example of pretesting: Nyamathi and colleagues (2005) studied the predictors of perceived health status in a sample of 415 homeless adults with tuberculosis. The study involved collecting an extensive array of data via self-reports. All of the instruments had been previously tested with homeless people, and many were pretested in group settings to determine clarity and sensitivity to the population.

Developing Data Collection Forms and Procedures

After the instrument package is finalized, researchers face several administrative tasks, such as the development of various forms (e.g., screening forms to assess eligibility, informed consent forms, records of attempted contacts with participants, logs for recording the receipt of data). It is prudent to design forms that are attractively formatted, legible, and inviting to use, especially if they are to be used by participants themselves. Care should also be taken to design forms to ensure confidentiality. For example, identifying information (e.g., names, addresses) is often recorded on a page that can be detached and kept separate from other data.

TIP: Whenever possible, try to avoid reinventing the wheel. It is inefficient and unnecessary to start from scratch—not only in developing instruments but also in creating forms, training materials, and so on. Ask seasoned researchers if they have materials you could borrow or adapt.

In most quantitative studies, researchers develop **data collection protocols** that spell out procedures to be used in data collection. These protocols describe such things as the following:

- Conditions that must be met for collecting the data (e.g., Can others be present at the time of data collection? Where must data collection occur?)
- Specific procedures for collecting the data, including requirements for sequencing multiple instruments and recording information
- Information to provide participants who ask routine questions about the study (i.e., answers to FAQs). Examples include the following: How will the information from this study be used? How did you get my name, and why are you asking me? How long will this take? Who will have access to this information? Can I see the study results? Whom can I contact if I have a complaint? Will I be paid or reimbursed for expenses?
- Procedures to follow in the event that a participant becomes distraught or disoriented, or for any other reason cannot complete the data collection

Researchers also need to decide how to actually gather, record, and manage their data. Technological advances continue to offer new options. As noted in Chapter 11, survey researchers are increasingly using sophisticated computer programs to facilitate collecting, recording, and encoding self-report data (e.g., CATI, CAPI). The Internet is being used to gather data from geographically dispersed populations. Personal digital assistants (PDAs) and audio-enhanced PDAs are also beginning to play a role. Courtney and Craven (2005) and Guadagno and colleagues (2004) offer some suggestions about new technology and data collection.

TIP: Document all major activities and decisions as you develop and implement your data collection plan, and save your documentation. You may need the information later when you write your research report, request funding for a follow-up study, or help other researchers.

STRUCTURED SELF-REPORT INSTRUMENTS

The most widely used data collection method by nurse researchers is structured self-report, which involves a formal, written instrument. The instrument is an **interview schedule** when questions are asked orally in face-to-face or telephone interviews. It is called a **questionnaire** or an SAQ (self-administered questionnaire) when respondents complete the instrument themselves, either in a paper-and-pencil format or on a computer. Researchers sometimes embed an SAQ into an interview schedule, with interviewers asking some questions orally but respondents answering others in writing. This section discusses the development and administration of structured self-report instruments.

Types of Structured Questions

Structured instruments consist of a set of questions (often called **items**) in which the wording of both the questions and, in most cases, *response alternatives* is predetermined. When structured instruments are used, people are asked to respond to the same questions, in the same order, and with the same set of response options. In developing structured instruments, much effort must be devoted to the content, form, and wording of questions.

Open and Closed Questions

Structured instruments vary in degree of structure through different combinations of open-ended and closed-ended questions. **Open-ended questions** allow people to respond in their own words, in narrative fashion. The question, "What was your biggest challenge after your surgery?" is an example of an open-ended question. In questionnaires, respondents are asked to give a written reply to open-ended items and so adequate space must be provided to permit a full response. Interviewers are expected to quote oral responses verbatim or as closely as possible.

Closed-ended (or **fixed-alternative**) **questions** offer response options, from which respondents

must choose the one that most closely matches the appropriate answer. The alternatives may range from a simple *yes* or *no* ("Have you smoked a cigarette within the past 24 hours?") to complex expressions of opinion or behavior.

Both open- and closed-ended questions have certain strengths and weaknesses. Good closed-ended items are often difficult to construct but easy to administer and, especially, to analyze. With closed-ended questions, researchers need only tabulate the number of responses to each alternative to gain descriptive information. The analysis of open-ended items, by contrast, is more difficult and time-consuming. The usual procedure is to develop categories and code open-ended responses into the categories. That is, researchers essentially transform open-ended responses to fixed categories in a post hoc fashion so that tabulations can be made.

Closed-ended items are more efficient than open-ended questions in that respondents can complete more closed- than open-ended questions in a given amount of time. In questionnaires, participants may be less willing to compose written responses than to check off appropriate alternatives. Closed-ended items are also preferred if respondents are unable to express themselves well verbally. Furthermore, some questions are less objectionable in closed form than in open form. Take the following example:

1. What was your family's total annual income last year?
2. In what range was your family's total annual income last year?
   - ❑ 1. Under $25,000,
   - ❑ 2. $25,000 to $49,999,
   - ❑ 3. $50,000 to $74,999,
   - ❑ 4. $75,000 to $99,999, or
   - ❑ 5. $100,000 or more

The second question gives respondents a greater measure of privacy than the open-ended question, and is less likely to go unanswered.

A major drawback of closed-ended questions is the possibility of omitting key responses. Such omissions can lead to inadequate understanding of the issues or to outright bias if respondents choose an alternative that misrepresents their position. Another objection to closed-ended items is that they tend to be superficial. Open-ended questions allow for a richer and fuller perspective on a topic, if respondents are verbally expressive and cooperative. Some of this richness may be lost when researchers tabulate answers they have categorized, but direct excerpts from open-ended responses can be valuable in imparting the flavor of the replies. Finally, some people may object to being forced into choosing from response options that do not reflect their opinions well. Open-ended questions give freedom to respondents and, therefore, offer the possibility of spontaneity and elaboration.

Decisions about the mix of open- and closed-ended questions is based on such considerations as the sensitivity of the questions, respondents' verbal ability, the amount of time available, and the amount of prior research on the topic. Combinations of both types can be used to offset the strengths and weaknesses of each. Questionnaires typically use closed-ended questions primarily, to minimize respondents' writing burden. Interview schedules, on the other hand, tend to be more variable in their mixture of these two question types.

Specific Types of Closed-Ended Questions

The analytic advantages of closed-ended questions are often compelling. Various types of closed-ended questions, illustrated in Table 13.1, are described here. Question types can be intermixed within a structured instrument.

- **Dichotomous questions** require respondents to make a choice between two response alternatives, such as yes/no or male/female. Dichotomous questions are especially appropriate for gathering factual information.
- **Multiple-choice questions** offer three or more response alternatives. Graded alternatives are preferable to dichotomous items for opinion or attitude questions because researchers get more information (intensity as well as direction of opinion) and respondents can express a range of views. Multiple-choice questions typically offer three to seven options.
- **Rank-order questions** ask respondents to rank target concepts along a continuum, such as most to least important. Respondents are asked

TABLE 13.1 Examples of Closed-Ended Questions

QUESTION TYPE	EXAMPLE
1. Dichotomous question	Have you ever been pregnant? 1. Yes 2. No
2. Multiple-choice question	How important is it to you to avoid a pregnancy at this time? 1. Extremely important 2. Very important 3. Somewhat important 4. Not important
3. Rank-order question	People value different things in life. Below is a list of things that many people value. Please indicate their order of importance to you by placing a "1" beside the most important, "2" beside the second-most important, and so on. ____ Career achievement/work ____ Family relationships ____ Friendships, social interactions ____ Health ____ Money ____ Religion
4. Forced-choice question	Which statement most closely represents your point of view? 1. What happens to me is my own doing. 2. Sometimes I feel I don't have enough control over my life.
5. Rating question	On a scale from 0 to 10, where 0 means "extremely dissatisfied" and 10 means "extremely satisfied," how satisfied were you with the nursing care you received during your hospitalization? 0 1 2 3 4 5 6 7 8 9 10 Extremely dissatisfied ... Extremely satisfied

to assign a 1 to the concept that is most important, a 2 to the concept that is second in importance, and so on. Rank-order questions can be useful, but respondents sometimes misunderstand them so good instructions and an example may be needed. Rank-order questions should involve 10 or fewer rankings.

- **Forced-choice questions** require respondents to choose between two statements that represent polar positions or characteristics.
- **Rating questions** ask respondents to evaluate something along an ordered dimension. Rating questions are typically on a **bipolar scale**, with end points specifying opposite extremes on a continuum. The end points and sometimes intermediary points along the scale are verbally labeled. The number of gradations or points along the scale can vary but often is an odd number, such as 7, 9, or 11, to allow for a neutral midpoint. (In the example in Table 13.1, the rating question has 11 points, numbered 0 to 10.)
- **Checklists** include several questions with the same response format. A checklist is a

The next question is about things that may have happened to you personally. Please indicate how recently, if ever, these things happened to you:

	Yes, within past 12 months	Yes, 2–3 years ago	Yes, more than 3 years ago	No, never
a. Has someone ever yelled at you all the time or put you down on purpose?	1	2	3	4
b. Has someone ever tried to control your every move?	1	2	3	4
c. Has someone ever threatened you with physical harm?	1	2	3	4
d. Has someone ever hit, slapped, kicked, or physically harmed you?	1	2	3	4

FIGURE 13.1 Example of a checklist.

two-dimensional arrangement in which a series of questions is listed along one dimension (usually vertically) and response alternatives are listed along the other. Checklists are relatively efficient and easy to understand, but because they are difficult to read orally, they are used more frequently in SAQs than in interviews. Figure 13.1 presents an example of a checklist.

- **Visual analog scales** (**VAS**) are used to measure subjective experiences, such as pain, fatigue, and dyspnea. The VAS is a straight line, the end anchors of which are labeled as the extreme limits of the sensation or feeling being measured. People are asked to mark a point on the line corresponding to the amount of sensation experienced. Traditionally, the VAS line is 100 mm in length, which facilitates the derivation of a score from 0 to 100 through simple measurement of the distance from one end of the scale to the person's mark on the line. An example of a VAS is shown in Figure 13.2.

In certain situations, researchers collect information about activities and dates, sometimes using an **event history calendar** (Martyn & Belli, 2002). Such calendars are matrixes that plot time on one dimension (usually the horizontal dimension) and the events or activities on the other. The person recording the data (either the participant or an interviewer) draws lines to indicate the stop and start dates of the specified events or behaviors. Event history calendars are especially useful in collecting information about the occurrence and sequencing of events retrospectively. Data quality about past occurrences is enhanced because the calendar helps participants relate the timing of some events to the timing of others. An example of an event history calendar is included in the Toolkit section of the accompanying *Resource Manual.*

An alternative to collecting event history data retrospectively is to ask participants to maintain information in an ongoing structured **diary** over a specified time period. This approach is often used to collect quantitative information about sleeping, eating, or exercise behavior.

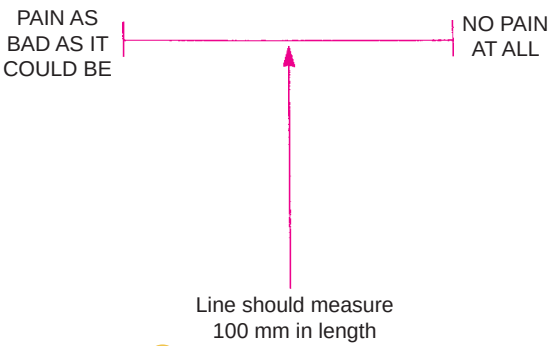


FIGURE 13.2 Example of a visual analog scale.

Example of a structured diary: Berger and colleagues (2009) examined the effect of menopausal status on sleep. Several sleep outcomes (e.g., total sleep time in minutes, number of awakenings, and minutes awake after sleep onset) were captured in daily diaries.

Composite Scales and Other Structured Self-Reports

Several special types of structured self-reports are used by nurse researchers. The most important are composite social-psychological scales that are often included in a questionnaire or interview package. A **scale** provides a numeric score to place respondents on a continuum with respect to an attribute, like a scale for measuring people's weight. Scales are used to discriminate quantitatively among people with different attitudes, fears, and needs. Scales are created by combining several closed-ended items into a composite score. Many sophisticated scaling techniques have been developed, but only two are discussed in this book.* We also briefly describe cognitive and neurologic tests, vignettes, and Q sorts.

Likert Scales

The most widely used scaling technique is the **Likert scale**, named after the psychologist Rensis Likert. A Likert scale consists of several declarative items that express a viewpoint on a topic. Respondents typically are asked to indicate the degree to which they agree or disagree with the opinion expressed by the statement.

Table 13.2 illustrates a six-item Likert-type scale for measuring attitudes toward condom use. Likert scales often include 10 or more statements; the example in Table 13.2 is shown only to illustrate key features. After respondents complete a Likert scale, their responses are scored. Typically, agreement with positively worded statements and disagreement with negatively worded ones are assigned higher scores. (See Chapter 15, however, for a discussion of problems in including both positive and negative items on a scale). The first statement in Table 13.2 is positively worded; agreement indicates a favorable attitude toward condom use. Thus, a higher score would be assigned to those agreeing with this statement than to those disagreeing with it. With five response alternatives, a score of 5 would be given to those strongly agreeing, 4 to those agreeing, and so forth. The responses of two hypothetical respondents are shown by a check or an X, and their scores are shown in far right columns. Person 1, who agreed with the first statement, has a score of 4, whereas person 2, who strongly disagreed, has a score of 1. The second statement is negatively worded, and so scoring is reversed—a 1 is assigned to those who strongly agree, and so on. This reversal is needed so that a high score consistently reflects positive attitudes toward condoms. A person's total score is computed by adding together individual item scores. Such scales are often called **summated rating scales** because of this feature. The total scores of both respondents are shown at the bottom of Table 13.2. The scores reflect a much more positive attitude toward condoms on the part of person 1 than person 2 does.

The summation feature of such scales makes it possible to make fine discriminations among people with different points of view. A single question allows people to be put into only five categories. A six-item scale, such as the one in Table 13.2, permits finer gradation—from a minimum possible score of 6 (6×1) to a maximum possible score of 30 (6×5).

Summated rating scales can be used to measure a wide array of attributes. In such cases, the bipolar scale may not be an agree/disagree continuum, but might be always true/never true, very likely/very unlikely, and so on. Constructing a good Likert-type scale requires considerable skill and work. Chapter 15 describes the steps involved in developing and testing such scales.

*Other scaling procedures include **ratio scaling, magnitude estimation scaling, multidimensional scaling**, and **multiple scalogram analysis**. Textbooks on psychological scaling and psychometric procedures should be consulted for more information about these scaling strategies.

TABLE 13.2 Example of a Likert Scale

DIRECTION OF SCORING*	ITEM	RESPONSES†					SCORE	
		SA	A	?	D	SD	Person 1 (✔)	Person 2 (×)
+	1. Using a condom shows you care about your partner.		✔			×	4	1
−	2. My partner would be angry if I talked about using condoms.			×		✔	5	3
−	3. I wouldn't enjoy sex as much if my partner and I used condoms.		×		✔		4	2
+	4. Condoms are a good protection against AIDS and other sexually transmitted diseases.			✔	×		3	2
+	5. My partner would respect me if I insisted on using condoms.	✔				×	5	1
−	6. I would be too embarrassed to ask my partner about using a condom.		×			✔	5	2
	Total score						26	11

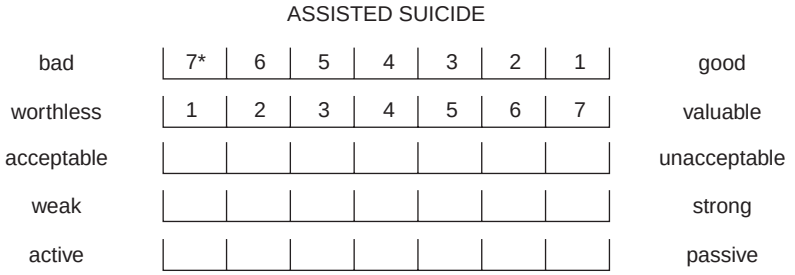
*Researchers would not indicate the direction of scoring on a Likert scale administered to study participants. The scoring direction is indicated in this table for illustrative purposes only.
†SA, strongly agree; A, agree; ?, uncertain; D, disagree; SD, strongly disagree.

Example of a summated rating scale: Lynn and colleagues (2009) developed a Likert-type scale to measure satisfaction in nursing. Examples of statements include the following: "Nurses on my unit enjoy working together" and "I enjoy being responsible for the welfare of my patients." Responses are on a 4-point scale, without a neutral response option.

Semantic Differential Scales

Another technique for measuring attitudes is the **semantic differential** (SD). With the SD, respondents are asked to rate concepts (e.g., dieting, exercise) on a series of *bipolar adjectives*, such as good/bad, effective/ineffective, important/unimportant. Respondents place a check at the appropriate point on a seven-point scale that extends from one extreme of the dimension to the other. Figure 13.3 shows an abbreviated example of the format for an SD for the concept *Assisted Suicide*.

SDs are flexible and easy to construct, and the concept being rated can be virtually anything—a person, concept, controversial issue, and so on. Scoring for SD responses is similar to that for



*The score values would not be printed on the form administered to actual participants. The numbers are presented here solely for the purpose of illustrating how semantic differentials are scored.

FIGURE 13.3 Example of a semantic differential.

Likert scales. Scores from 1 to 7 are assigned to each bipolar scale response, with higher scores generally associated with the positively worded adjective. Responses are then summed across the bipolar scales to yield a total score.

Researchers can be creative in their choice of bipolar scales, but the adjective pairs should be appropriate for the concepts. The adjective pair large/small for the SD in Figure 13.3 would not make much sense. Another consideration in selecting adjective pairs is the extent to which the adjectives measure the same dimension of the concept. Research with SD scales suggests that adjective pairs tend to cluster along three independent dimensions: evaluation, potency, and activity. Evaluative adjectives, such as effective/ineffective or good/bad are especially important. Potency adjectives include strong/weak and large/small, and examples of activity adjectives are active/passive and fast/slow. These three dimensions need to be scored separately because people's *evaluative* ratings of a concept are independent of their *activity* or *potency* ratings. Researchers must decide how many SD dimensions to include.

Example of a study using an SD: Rempusheski and O'Hara (2005) developed a semantic differential scale, the Grandparent Perceptions of Family Scale (GPFS). Respondents rate stimuli (e.g., "How I view my grandchild") with regard to 22 bipolar adjective pairs. Three adjective pairs were in the action subscale (e.g., active/passive), 11 were in the evaluative subscale (e.g., happy/sad), and 8 were in the potency subscale (e.g., emotionally strong/emotionally weak).

TIP: Most nurse researchers use existing scales rather than developing their own. Resources for locating existing scales include Strickland and Dilorio, 2003; Frank-Stromberg and Olsen, 2004; and Waltz and colleagues, 2010. Also, some helpful websites are included in the Toolkit. Another place to look for existing instruments is in the Health and Psychosocial Instruments (HaPI) database.

Cognitive and Neuropsychological Tests

Nurse researchers sometimes assess study participants' cognitive skills. There are several different types of **cognitive tests**. For example, *intelligence tests* evaluate a person's global ability to perceive relationships and solve problems and *aptitude tests* measure a person's potential for achievement. Some tests have been developed for individual (one-on-one) administration, whereas others have been developed for group use. Individual tests, such as the Stanford-Binet I.Q. test, must be administered by a person with special training. Nurse researchers are especially likely to use ability tests in studies of high-risk groups, such as low-birth-weight children.

Some cognitive tests are specially designed to assess neuropsychological functioning among people with potential cognitive impairments, such as the Mini-Mental Status Examination (MMSE). These tests capture varying types of competence, such as the ability to concentrate and the ability to remember. Nurses have used such tests extensively in studies of elderly patients and patients with Alzheimer's disease. Good sources for learning more about ability tests are the books by Urbina (2004) and the Buros Institute (2007).

Example of a study assessing neuropsychological function: Alpert and colleagues (2009) did a pilot study to evaluate the effect of jazz dance instruction on balance, cognition, and mood in community-dwelling older women. Cognitive outcomes were measured using the MMSE.

Q Sorts

In a **Q sort**, participants are presented with a set of cards on which words or phrases are written. Participants are told to sort the cards along a specified bipolar dimension, such as most important/least important. Typically, there are between 50 and 100 cards to be sorted into 9 or 11 piles, with the number of cards to be placed in each pile predetermined by the researcher (e.g., 2 cards in piles 1 and 9, 4 cards in piles 2 and 8, and so on). It is difficult to achieve reliable results with fewer than 50 cards, but the task becomes tedious and difficult with more than 100.

The sorting instructions and objects to be sorted in a Q sort can vary. For example, attitudes can be studied by writing attitudinal statements on the cards and asking participants to sort them on a continuum from "totally disagree" to "totally agree." Or, patients could be asked to rate nursing behaviors on a continuum from least helpful to most helpful.

Q sorts are versatile and can be applied to a wide variety of problems. Requiring people to place a predetermined number of cards in each pile can reduce biases that are common in Likert scales. On the other hand, it is difficult and time-consuming to administer Q sorts to a large sample of people. Some critics argue that the forced distribution of cards according to researchers' specifications is artificial and excludes information about how participants would ordinarily distribute their responses. Another issue is that Q sorts cannot be incorporated into mailed or Internet questionnaires or administered in telephone interviews. The paper by Akhtar-Danesh and colleagues (2008) provides more information about Q sorts.

Example of a Q sort: Akhtar-Danesh and colleagues (2008) used a 43-card Q sort to examine nurse faculty perceptions of simulation use in nursing education. Statements were sorted into 9 piles on an agree/disagree continuum. An example of a statement in the card sort is: "It's a scheduling nightmare."

Vignettes

Another self-report approach involves the use of **vignettes**, which are brief case reports or descriptions of events to which respondents are asked to react. The descriptions, which can either be fictitious or based on fact, are structured to elicit information about respondents' perceptions of some phenomenon or their projected actions. The vignettes are usually written narrative descriptions, but researchers have also used videotaped vignettes. The questions that follow the vignettes can be open-ended (e.g., How would you describe this patients' level of confusion?) or closed-ended (e.g., Rate how confused you think this patient is on a 7-point scale). Usually 3 to 5 vignettes are included in an instrument.

Sometimes the underlying purpose of vignette studies is not revealed to participants, especially if the technique is used as an indirect measure of prejudices or stereotypes using embedded descriptors, as in the following example.

Example of vignettes: Griffin and colleagues (2007) distributed vignette packets describing three hospitalized children to a national sample of pediatric nurses to explore whether pain management decisions were affected by children's characteristics. Three vignettes described children in pain: one described either a boy or a girl, another described a white or African American child, and the third described a physically attractive or unattractive child. Nurses answered questions about pain treatments they would use without being aware that the child characteristics had been experimentally varied.

Vignettes are an economical means of eliciting information about how people might behave in situations that would be difficult to observe in daily life. Vignettes can be incorporated into questionnaires, and are, therefore, an inexpensive data collection strategy. Also, vignettes often represent an interesting task for participants. The principal problem with vignettes concerns the validity of responses. If respondents describe how they would act in a situation portrayed in the vignette, how accurate is that description of their actual behavior? Thus, although the use of vignettes can be profitable, potential biases should be taken into account in interpreting results.

TIP: Some methods described in this chapter might be appealing because they are unusual and may seem like a creative approach to collecting data. However, the prime considerations in selecting a data collection method should always be the conceptual congruence between the method and the targeted constructs, and the quality of data that the method yields.

Questionnaires Versus Interviews

In developing their data collection plans, researchers need to decide whether to collect data through interviews or questionnaires. Each method has advantages and disadvantages.

Advantages of Questionnaires

Self-administered questionnaires, which can be distributed in person, by mail, or over the Internet, offer some advantages. The strengths of questionnaires include the following:

- *Cost*. Questionnaires, relative to interviews, are much less costly. Distributing questionnaires to groups (e.g., nursing home residents) is inexpensive and expedient. And, with a fixed amount of funds or time, a larger and more geographically diverse sample can be obtained with mailed or Internet questionnaires than with interviews.
- *Anonymity*. Unlike interviews, questionnaires offer the possibility of complete anonymity. A guarantee of anonymity can be crucial in obtaining candid responses, particularly if questions are sensitive. Anonymous questionnaires often result in a higher proportion of socially unacceptable responses (i.e., responses that place respondents in an unfavorable light) than interviews.
- *Interviewer bias*. The absence of an interviewer ensures that there will be no interviewer bias. Interviewers ideally are neutral agents through whom questions and answers are passed. Studies have shown, however, that this ideal is difficult to achieve. Respondents and interviewers interact as humans, and this interaction can affect responses.

Internet surveys are especially economical and can sometimes yield a dataset directly amenable to analysis, without requiring someone to enter data onto a file (the same is also true for CAPI and CATI interviews). Internet surveys also provide opportunities for providing participants with customized feedback and for prompts that can minimize missing responses.

Advantages of Interviews

It is true that interviews are costly, prevent anonymity, and bear the risk of interviewer bias. Nevertheless, interviews are considered superior to questionnaires for most research purposes because of the following advantages:

- *Response rates*. Response rates tend to be high in face-to-face interviews. People are less likely to refuse to talk to an interviewer who directly solicits their cooperation than to ignore a questionnaire or email. A well-designed and properly conducted interview study normally achieves response rates in the vicinity of 80% to 90%, whereas mailed and Internet questionnaires typically achieve response rates of less than 50%. Because nonresponse is not random, low response rates can introduce serious biases. (However, if questionnaires are personally distributed in a particular setting—e.g., patients in a cardiac rehabilitation program—reasonably good response rates often can be achieved.)

TIP: MacDonald and colleagues (2009) have offered useful advice for addressing nonresponse in mailed surveys. Several suggestions are useful for minimizing nonresponse in collecting any type of self-report data. An additional useful resource is a meta-analysis of strategies to increase response to mailed and electronic surveys by Edwards and colleagues (2009).

- *Audience*. Many people cannot fill out a questionnaire. Examples include young children and blind, elderly, illiterate, or uneducated individuals. Interviews, on the other hand, are feasible with most people. For Internet questionnaires, a particularly important drawback is that not everyone has access to computers or uses them regularly.

- *Clarity*. Interviews offer some protection against ambiguous or confusing questions. Interviewers can assess whether questions have been misunderstood and provide clarification. With questionnaires, misinterpreted questions can go undetected.
- *Depth of questioning*. Information obtained from questionnaires tends to be more superficial than from interviews, largely because questionnaires usually contain mostly closed-ended items. Open-ended questions are avoided in questionnaires because most people dislike having to compose a reply. Furthermore, interviewers can enhance the quality of self-report data through *probing*, a topic we discuss later in this chapter.
- *Missing information*. Respondents are less likely to give "don't know" responses or to leave a question unanswered in an interview than on a questionnaire.
- *Order of questions*. In an interview, researchers have control over question ordering. Questionnaire respondents can skip around from one section to another. Sometimes a different ordering of questions from the one intended could bias responses.
- *Sample control*. Interviewers know whether the people being interviewed are the intended respondents. People who receive questionnaires, by contrast, can pass the instrument on to a friend or relative, and this can change the sample composition. Internet surveys are especially vulnerable to the risk that people not targeted by researchers will respond, unless there are password protections.
- *Supplementary data*. Face-to-face interviews can yield additional data through observation. Interviewers can observe and assess respondents' level of understanding, degree of cooperativeness, social class, and so forth. Such information can be useful in interpreting responses.

Many advantages of face-to-face interviews also apply to telephone interviews. Long or detailed interviews or ones with sensitive questions are not well suited to telephone administration, but for relatively brief instruments, telephone interviews are economical and tend to yield a higher response rate than mailed or Internet questionnaires.

Designing Structured Self-Report Instruments

Assembling a high-quality structured self-report instrument is demanding. To design useful, accurate instruments, researchers must carefully analyze the research requirements and attend to minute details. The steps for developing structured self-report instruments follow closely the ones we outlined earlier in the chapter, but a few additional considerations should be mentioned.

Related constructs should be clustered into separate **modules** or areas of questioning. For example, an interview schedule may consist of a module on demographic information, another on health symptoms, a third on stressful life events, and a fourth on health-promoting activities. Thought needs to be given to sequencing modules, and questions within modules, to arrive at an order that is psychologically meaningful and encourages candor. The schedule should begin with questions that are interesting, motivating, and not too sensitive. The instrument also needs to be arranged to minimize bias because early questions sometimes influence responses to subsequent ones. Whenever both general and specific questions about a topic are included, general questions should be placed first to avoid "coaching."

Instruments should be prefaced by introductory comments about the nature and purpose of the study. In interviews, introductory information would be communicated by the interviewer, who would typically follow a script. In questionnaires, the introduction usually takes the form of an accompanying **cover letter**. The introduction should be carefully constructed because it is the first point of contact with potential respondents. An example of a cover letter for a mailed questionnaire is presented in Figure 13.4. (This cover letter is included in the Toolkit for you to use and adapt.)

When a first draft of the instrument is in reasonably good order, it should be reviewed by experts in questionnaire construction, by substantive content

Dear Community Resident:

We are conducting a study to examine how men who are approaching retirement age (55 to 65 years old) feel about various issues relating to their healthcare. This study, which is sponsored by the National Institutes of Health, will enable healthcare providers to better meet the needs of men in your age group. Would you please assist us in this study by completing the enclosed questionnaire? Your opinions and experiences are very important to us and are needed to give an accurate picture of the health-related needs of men in the Capital District.

Your name was selected at random from a list of residents in your community. The questionnaire is completely anonymous, so you are not asked to put your name on it or identify yourself in any way. We hope, therefore, that you will feel comfortable giving your honest opinions. If you prefer not to answer any particular question, feel free to leave it blank. Please *do* answer questions if you can, though. If you have any comments or concerns about any questions, just write your comments in the margin of the questionnaire or feel free to contact me by email (dfp1@yahoo.com) or by phone (518-587-3994).

A postage-paid return envelope is enclosed for your convenience. Please take a few minutes to complete and return the questionnaire to us—it should only take about 15 to 20 minutes of your time. **In appreciation for your cooperation, you will be entered into a raffle to win a $250 American Express gift certificate.** Simply return the self-addressed, stamped postcard separately from the questionnaire. To be included in the raffle, your questionnaire must be returned to us by July 10. The raffle winner will be notified by July 17.

Your participation in the study is completely voluntary. By returning your study booklet, you will be granting your consent to participate in the study. Thank you in advance for your assistance.

FIGURE 13.4 Example of a cover letter.

area specialists, and by someone capable of detecting technical problems, such as spelling mistakes, grammatical errors, and so forth. When these various people have provided feedback, a revised version of the instrument can be pretested. The pretest should be administered to a small sample of individuals (usually 10 to 20) who are similar to actual participants.

In the remainder of this section, we offer some specific suggestions for designing high-quality self-report instruments. Additional guidance is offered in the classic book by Fowler (1995) and by Bradburn and colleagues (2004).

Tips for Wording Questions

We all are accustomed to asking questions, but the proper phrasing of questions for a study is not easy. In wording their questions, researchers should keep four important considerations in mind.

1. *Clarity*. Questions should be worded clearly and unambiguously. This is usually easier said than done. Respondents do not always have the same mind-set as the researchers.
2. *Ability of respondents to give information*. Researchers need to consider whether respondents can be expected to understand the question or are qualified to provide meaningful information.
3. *Bias*. Questions should be worded in a manner that will minimize the risk of response biases.
4. *Sensitivity*. Researchers should strive to be courteous, considerate, and sensitive to respondents' needs and circumstances, especially when asking questions of a private nature.

Here are some specific suggestions with regard to these four considerations (additional guidance on wording items for composite scales is provided in Chapter 15):

- Clarify in your own mind the information you are seeking. The question, "When do you usually

eat your evening meal?" might elicit such responses as "around 6 pm," "when my son gets home from soccer practice," or "when I feel like cooking." The question itself contains no words that are difficult, but the question is unclear because the researcher's intent is not apparent.

- Avoid jargon or technical terms (e.g., parity) if lay terms (e.g., number of children) are equally appropriate. Use words that are simple enough for the *least* educated respondents in the sample. Don't assume that even nurses have extensive knowledge on all aspects of nursing and medical terminology.
- Do not assume that respondents will be aware of, or informed about, issues in which you are interested. Furthermore, avoid giving the impression that they *ought* to be informed. Questions on complex issues sometimes can be worded in such a way that respondents will be comfortable admitting ignorance (e.g., "Many people have not had a chance to learn much about factors that increase the risk of diabetes. Do you happen to know of any contributing factors?") Another approach is to preface a question by a short explanation about terminology or issues.
- Avoid leading questions that suggest a particular answer. A question such as, "Do you agree that nurse-midwives play an indispensable role in the health team?" is not neutral.
- State a range of alternatives within the question itself when possible. For instance, the question, "Do you prefer to get up early in the morning on weekends?" is more suggestive of the "right" answer than "Do you prefer to get up early in the morning or to sleep late on weekends?"
- For questions that deal with controversial topics or socially unacceptable behavior (e.g., excessive drinking, noncompliance with medical regimens), closed-ended questions may be preferred. It is easier to check off having engaged in socially disapproved actions than to verbalize those actions in response to open-ended questions. Moreover, when controversial behaviors are presented as options, respondents are more likely to believe that their behavior is not unique, and admissions of such behavior become less difficult.
- Impersonal wording of questions is sometimes useful in encouraging honesty. To illustrate this point, compare these two statements with which respondents might be asked to agree or disagree: (1) "I am dissatisfied with the nursing care I received during my hospitalization" and (2) "The quality of nursing care in this hospital is unsatisfactory." A respondent might feel more comfortable admitting dissatisfaction with nursing care in the less personally worded second question.

Tips for Preparing Response Alternatives

If closed-ended questions are used, researchers also need to develop response alternatives. Below are some suggestions for preparing them.

- Responses options should cover all significant alternatives. If respondents are forced to choose from options provided by researchers, they should feel comfortable with the available options. As a precaution, researchers often have as a response option a phrase such as "Other—please specify."
- Alternatives should be mutually exclusive. The following categories for a question on a person's age are *not* mutually exclusive: 30 years or younger, 30 to 50 years, or 50 years or older. People who are exactly 30 or 50 would qualify for two categories.
- There should be a rationale for ordering alternatives. Options often can be placed in order of decreasing or increasing favorability, agreement, or intensity. When options have no "natural" order, alphabetic ordering of the alternatives can avoid leading respondents to a particular response (e.g., see the rank order question in Table 13.1).
- Response alternatives should be brief. One sentence or phrase for each option is usually sufficient to express a concept. Response alternatives should be about equal in length.

Tips for Formatting an Instrument

The appearance and layout of an instrument may seem a matter of minor administrative importance. Yet, a poorly designed format can have substantive consequences if respondents (or interviewers) become confused, miss questions, or answer

questions they should have omitted. The format is especially important in questionnaires because respondents cannot usually ask for help. The following suggestions may be helpful in laying out an instrument:

- Do not compress too many questions into too small a space. An extra page of questions is better than a form that appears dense and confusing and that provides inadequate space for responses to open-ended questions.
- Set off the response options from the question or stem. Response alternatives are usually aligned vertically (Table 13.1). In questionnaires, respondents can be asked either to circle their answer or to check the appropriate box.
- Give special care to formatting **filter questions**, which are designed to route respondents through different sets of questions depending on their responses. In interview schedules, the typical procedure is to use **skip patterns** that instruct interviewers to skip to a specific question (e.g., SKIP TO Q10). In SAQs, skip instructions may be confusing. It is usually better to put questions appropriate to a subset of respondents apart from the main series of questions, as illustrated in Box 13.1, part B. An important advantage of CAPI, CATI, audio-CASI, and some Internet surveys is that skip patterns are built into the computer program, leaving no room for human error.
- Avoid forcing all respondents to go through inapplicable questions in an SAQ. That is, question 2 in Box 13.1 part B could have been worded as follows: "If you are a member of the American Nurses Association, for how long have you been a member?" Nonmembers may not be sure how to handle this question and may be annoyed at having to read through irrelevant material.

Administering Structured Self-Report Instruments

Administering interview schedules and questionnaires involves different considerations and requires different skills.

BOX 13.1 Examples of Formats for a Filter Question

A. *Interview Format*

1. Are you currently a member of the American Nurses Association?
   - ❑ 1. Yes
   - ❑ 2. No (SKIP TO Q3)
2. For how many years have you been a member?
   _____YEARS
3. Do you subscribe to any nursing journals?
   - ❑ 1. Yes
   - ❑ 2. No

B. *Questionnaire Format*

1. Are you currently a member of the American Nurses Association?
   - ❑ 1. Yes
   - ❑ 2. No

2. If yes: For how many years have you been a member?
   _____YEARS

3. Do you subscribe to any nursing journals?
   - ❑ 1. Yes
   - ❑ 2. No

Collecting Interview Data

The quality of interview data relies heavily on interviewer proficiency. Interviewers for large survey organizations receive extensive general training in addition to specific training for individual studies. Although we cannot in this introductory book cover all the principles of good interviewing, we can identify some major issues. Additional guidance can be found in the classic handbook by Fowler and Mangione (1990).

A primary task of interviewers is to put respondents at ease so that they will feel comfortable in expressing opinions honestly. Respondents' reactions to interviewers can affect their level of cooperation. Interviewers, therefore, should always be punctual (if an appointment has

been made), courteous, and friendly. Interviewers should strive to appear unbiased and to create an atmosphere that encourages candor. All opinions of respondents should be accepted as natural; interviewers should not express surprise, disapproval, or even approval.

With a structured interview schedule, interviewers should follow question wording precisely. Interviewers should not offer spontaneous explanations of what questions mean. Repetition of a question is usually adequate to dispel misunderstandings, especially if the instrument has been pretested. Interviewers should not read questions mechanically. A natural, conversational tone is essential in building rapport, and this tone is impossible to achieve if interviewers are not thoroughly familiar with the questions.

When closed-ended questions have lengthy or complex response alternatives, or when a series of questions has the same response options, interviewers should hand respondents a **show card** that lists the options. People cannot be expected to remember detailed unfamiliar material and may choose the last alternative if they cannot recall earlier ones. (Examples of show cards are included in the Toolkit in the *Resource Manual*.)

Interviewers record answers to closed-ended items by checking or circling the appropriate alternative, but responses to open-ended questions must be written out in full. Interviewers should not paraphrase or summarize respondents' replies.

Obtaining complete, relevant responses to questions is not always an easy matter. Respondents may reply to seemingly straightforward questions with partial answers. Some may say, "I don't know" to avoid giving their opinions on sensitive topics, or to stall while they think over the question. In such cases, the interviewers' job is to **probe**. The purpose of a probe is to elicit more useful information than respondents volunteered during their initial reply. A probe can take many forms: Sometimes it involves repeating the original question, and sometimes it is a long pause intended to communicate to respondents that they should continue. Frequently, it is necessary to encourage a more complete response to open-ended questions by a nondirective supplementary question, such as, "How is that?" Interviewers must be careful to use only *neutral* probes that do not influence the content of a response. Box 13.2 gives some examples of neutral, nondirective probes used by professional interviewers to get more complete responses to questions. The ability to probe well is perhaps the greatest test of an interviewer's skill. To know when to probe and how to select the best probes, interviewers must understand the purpose of each question. (The Toolkit for Chapter 14 has material relating to interviewer training that might be useful.)

BOX 13.2 Examples of Neutral, Nondirective Probes

- Is there anything else?
- Go on.
- Are there any other reasons?
- How do you mean?
- Could you please tell me more about that?
- Would you tell me what you have in mind?
- There are no right or wrong answers; I'd just like to get your thinking.
- Could you please explain that?
- Could you please give me an example?

Guidelines for telephone interviews are essentially the same as those for face-to-face interviews, but additional effort usually is required to build rapport over the telephone. In both cases, interviewers should strive to make the interview a pleasant and satisfying experience in which respondents are made to understand that the information they are providing is important.

Collecting Questionnaire Data through In-Person Distribution

Questionnaires can be distributed in various ways, including personal distribution, through the mail, and over the Internet. The most convenient procedure is to distribute questionnaires to a group of people who complete the instrument at

the same time. This approach has the obvious advantages of maximizing the number of completed questionnaires and allowing respondents to ask questions. Group administrations are often possible in educational settings and in some clinical situations.

Researchers can also hand out questionnaires to individual respondents. Personal contact has a positive effect on response rates, and researchers can answer questions. Individual distribution of questionnaires in clinical settings is often inexpensive and efficient and can yield a relatively high rate of response.

Example of personal distribution of questionnaires: Dirksen and colleagues (2009) explored the relationships between insomnia, depression, and distress in men with prostate cancer. Data were collected by means of questionnaires that were distributed by a research assistant to men receiving treatment in an outpatient ambulatory clinic.

Collecting Questionnaire Data through the Mail

For surveys of a broad population, questionnaires are often mailed. This approach is cost-effective for reaching geographically dispersed respondents, but it tends to yield low response rates. When only a subsample of respondents return their questionnaires, the risk of bias is high. With low response rates, researchers face the possibility that people who did not complete a questionnaire would have answered questions differently from those who did return it.

With response rates greater than 65%, the risk of bias may be relatively small, but lower response rates are the norm. Researchers should attempt to discover how representative respondents are, relative to the selected sample, in terms of basic demographic characteristics, such as age, gender, and race/ethnicity. This comparison may lead researchers to conclude that respondents and nonrespondents are sufficiently similar. When demographic differences are found, investigators can make inferences about the direction of biases.

Response rates can be affected by the manner in which the questionnaires are designed and mailed. The physical appearance of the questionnaire can influence its appeal, so thought should be given to instrument layout, quality and color of paper, and method of reproduction. The standard procedure for distributing mailed questionnaires is to include a stamped, addressed return envelope—without which, response rates will be seriously jeopardized.

TIP: People are more likely to complete a mailed questionnaire if they are encouraged to do so by someone whose name (or position) they recognize. If possible, include a letter of endorsement from someone visible (e.g., a hospital or government official), or write the cover letter on the stationery of a well-respected organization, such as a university.

Follow-up reminders are effective in achieving higher response rates for mailed (and Internet) questionnaires. This procedure involves additional mailings urging nonrespondents to complete and return their forms. Follow-up reminders are typically sent about 10 to 14 days after the initial mailing. Sometimes reminders simply involve a letter or postcard of encouragement to nonrespondents. It is preferable, however, to send a second copy of the questionnaire with the reminder letter because many nonrespondents will have misplaced or discarded the original. Telephone follow-ups can be even more successful, but are costly and time-consuming. With anonymous questionnaires, researchers may be unable to distinguish respondents and nonrespondents for the purpose of sending follow-up letters. In such a situation, the simplest procedure is to send out a follow-up reminder to the entire sample, thanking those who have already answered and asking others to cooperate. Dillman and colleagues (2009) offer excellent advice for achieving acceptable response rates in mailed and Internet surveys.

Example of mailed questionnaires: Kupferer and colleagues (2009) surveyed women who had discontinued hormone therapy with regard to their use of complementary and alternative medicine for vasomotor symptoms. Questionnaire packets and a postage-paid return envelope were mailed to a random sample of 2,250 women from a purchased mailing list. The response rate was 24%.

Collecting Questionnaire Data via the Internet

The Internet is an economical means of distributing questionnaires. Internet surveys appear to be a promising approach for accessing groups of people interested in specific topics. Internet distribution requires appropriate equipment and some technical skills, but there are a growing number of aids for doing such surveys.

Surveys can be administered through the Internet in several ways. One method is to design a questionnaire in a word processing program, as would be the case for mailed questionnaires. The file with the questionnaire is then attached to an email message and distributed to potential respondents. Respondents can complete the questionnaire and return it as an email attachment or print it and return it by mail or fax. This method may be problematic if respondents have trouble opening attachments or if they use a different word-processing program. Surveys sent via email also run the risk of not getting delivered to the intended party, either because email addresses have changed or because the email messages are blocked by Internet security filters. Blocks are especially common for messages with attachments.

Increasingly, researchers are collecting data through **web-based surveys**. This approach requires researchers to have a website on which the survey is placed or to use a service such as Survey Monkey (*http://www.surveymonkey.com/*). Respondents typically access the website by clicking on a hypertext link. For example, respondents may be invited to participate in the survey through an email message that includes the hyperlink to the survey, or they may be invited to participate when they enter a website related in content to the survey (e.g., the website of a cancer support organization).

Web-based forms are designed for online response, and some can be programmed to include interactive features. By having dynamic features, respondents can receive as well as give information—a feature that can increase motivation to participate. For example, respondents can be given information about their own responses (e.g., how they scored on a scale) or aggregated information about other participants. A major advantage of web-based surveys is that the data are directly amenable to analysis. They can, however, be more expensive than email surveys.

Example of a web-based survey: Sarna and colleagues (2009) conducted a web-based survey to obtain information from nurses in Magnet hospitals about their delivery of smoking cessation interventions. Respondents were solicited through the Chief Nursing Officers (CNOs) at 35 Magnet hospitals meeting inclusion criteria. CNOs were asked to communicate information about the survey web link to their nursing staff. The final response rate was 21%.

Internet surveys will undoubtedly abound in the years ahead. They tend to be economical and can reach a broad audience. However, samples are almost never representative, and response rates tend to be low—often even lower than mailed questionnaires. Several references are available to help researchers who wish to launch an Internet survey. For example, the books by Best and Krueger (2004), Dillman and colleagues (2009), and Fitzpatrick and Montgomery (2004) provide useful information. Weber and colleagues (2005) and Cantrell and Lupinacci (2007) offer guidance on web-based data collection and management.

Evaluation of Structured Self-Reports

Structured self-reports are a powerful data collection method. They are versatile and wide ranging, and yield information that can be readily analyzed statistically. Structured questions can be carefully worded and pretested. In an unstructured interview, by contrast, respondents may answer different questions, and there is no way to know whether question wording affected responses. On the other hand, the questions tend to be much more superficial than questions in unstructured interviews because most structured questions are closed-ended.

Structured self-reports are susceptible to the risk of various **response biases**—many of which are also possible in unstructured self-reports.

Respondents may give biased answers in reaction to the interviewers' behavior or appearance, for example. Perhaps the most pervasive problem is people's tendency to present a favorable image of themselves. **Social desirability response bias** refers to the tendency of some individuals to misrepresent themselves by giving answers that are congruent with prevailing social values. This problem is often difficult to combat. Subtle, indirect, and delicately worded questioning sometimes can help to minimize this response bias. The creation of a permissive atmosphere and provisions for anonymity also encourage frankness. In an interview situation, interviewer training is essential.

Some response biases, called **response sets**, are most commonly observed in composite scales. **Extreme responses** are a bias reflecting consistent selection of extreme alternatives (e.g., "strongly agree"). These extreme responses distort the findings because they do not necessarily signify the most intense feelings about the phenomenon under study, but rather capture a trait of the respondent. There is little a researcher can do to counteract this bias, but there are procedures for detecting it.

Some people have been found to agree with statements regardless of content. Such people are called **yea-sayers**, and the bias is known as the **acquiescence response set**. A less common problem is the opposite tendency for other individuals, called **naysayers**, to disagree with statements independently of question content.

Researchers who construct scales should attempt to eliminate or minimize response set biases. If an instrument or scale is being developed for general use by others, evidence should be gathered to demonstrate that the scale is sufficiently free from response biases to measure the critical variable. Users should consider such evidence in selecting existing scales.

STRUCTURED OBSERVATION

Structured observation is used to document specific behaviors, actions, and events. Structured observation involves using formal instruments and protocols that indicate what to observe, how long to observe it, and how to record information. The challenge of structured observation lies in the formulation of a system for accurately categorizing and recording observations.

In selecting behaviors, conversation, or attributes to be observed, researchers must decide what constitutes a *unit*. A *molar approach* entails observing large units of behavior and treating them as a whole. For example, an entire constellation of verbal and nonverbal behaviors might be construed as signaling confusion in nursing home residents. At the other extreme, a *molecular approach* uses small, specific behaviors or verbal segments as units. Each action, gesture, or phrase is treated as a separate entity. The molar approach is more susceptible to observer errors because of greater ambiguity in what is being observed. On the other hand, in reducing observations to concrete, specific elements, investigators may fail to understand how small elements work in concert in a behavior pattern. The choice of approach depends on the nature of the research problem.

Methods of Recording Structured Observations

Researchers recording structured observations typically use either a checklist or a rating scale. Both types of record-keeping instruments specify the behaviors or events to be observed and are designed to produce numeric information.

TIP: Compared with the abundance of books designed to provide guidance in developing self-report instruments, there are relatively few resources for researchers who want to design their own observational instruments, except if the focus of the observation is on interpersonal interactions (e.g., Kerig & Lindahl, 2001; Kerig & Baucom, 2004).

Category Systems and Checklists

Structured observation often involves constructing a category system to classify observed phenomena.

A **category system** represents an attempt to designate in a systematic fashion the qualitative behaviors and events transpiring in the observational setting.

Some category systems are constructed so that *all* observed behaviors within a specified domain (e.g., utterances) can be classified into one and only one category. In such an exhaustive system, the categories are mutually exclusive.

Example of exhaustive categories: Foreman and colleagues (2008) analyzed gender differences in the sleep–wake states of 97 preterm infants, who were videotaped in 4-hour segments. The infants' respirations, eye movements, facial expressions, muscle tone, and motor activity were used to classify their sleep–wake state, every 15 seconds, into one of four mutually exclusive categories: awake, drowsy, active sleep, and quiet sleep.

When observers use an exhaustive system—that is, when all behaviors of a certain type, such as verbal interaction, are observed and recorded—researchers must be careful to define categories so that observers know when one behavior ends and a new one begins. Another essential feature is that referent behaviors should be mutually exclusive, as in the previous example. The underlying assumption in using such a category system is that behaviors, events, or attributes that are allocated to a particular category are equivalent to every other behavior, event, or attribute in that same category.

A contrasting technique is to develop a system in which only particular types of behavior (which may or may not be manifested) are categorized. For example, if we were studying autistic children's aggressive behavior, we might develop such categories as "strikes another child," or "kicks or hits walls or floor." In such a category system, many behaviors—all the ones that are nonaggressive—would not be classified. Nonexhaustive systems are adequate for many purposes, but one risk is that resulting data might be difficult to interpret. Problems may arise if a large number of behaviors are not categorized or if long segments of the observation sessions do not involve the target behaviors. In such situations, investigators need to record the amount of time in which the target behaviors occurred, relative to the total time under observation.

Example of nonexhaustive categories: Liaw and colleagues (2006) studied changes in patterns of infants' distress at different phases of a routine tub bath in the NICU. The researchers developed a system to categorize behavioral signs of distress (jerks, tremors, grimaces, arching). Behaviors unrelated to distress were not categorized.

A critical requirement for a good category system is the careful definition of behaviors or characteristics to be observed. Each category must be explained in detail so that observers have relatively clear-cut criteria for identifying the occurrence of a specified phenomenon. Virtually all category systems require observers to make some inferences, to a greater or lesser degree.

Example of low observer inference: Johnston and colleagues (2008) studied the effects of kangaroo mother care on preterm infants' pain from a heel lance. They used the Premature Infant Pain Profile (PIPP) to measure pain. The PIPP includes both physiologic (e.g., heart rate) and behavioral indicators. Three facial actions (brow bulge, eye squeeze, and naso-labial furrow) are scored by observers. The coding system "provides a detailed, anatomically based, and objective description" (p. 4) of newborn behavior.

In this system, assuming that observers were properly trained, relatively little inference would be required to code facial actions. Other category systems, however, require more inference, as in the following example:

Example of moderately high observer inference: Uitterhoeve and colleagues (2008) videotaped oncology nurses interacting with actors playing the role of patients. The videotaped encounters were coded for nurses' responses to patients' cues. Nurses' responses were coded according to both function and form. Function, for example, involved coding whether the patient's cue was *explored, acknowledged* but not explored, or elicited a *distancing* response.

In such category systems, even when categories are defined in detail, a moderately heavy inferential

burden is placed on observers. The decision concerning degree of observer inference depends on a number of factors, including the research purpose and the observers' skills. Beginning researchers are advised to construct or use category systems that require low to moderate inference.

Category systems are used to construct a **checklist**, which is the instrument observers use to record observed phenomena. The checklist is usually formatted with the list of behaviors or events from the category system on the left and space for tallying the frequency or duration of occurrence of behaviors on the right. With nonexhaustive category systems, categories of behaviors that may or may not be manifested by participants are listed on the checklist. The observer's tasks are to watch for instances of these behaviors and to record their occurrence.

With exhaustive checklists, the observers' task is to place all behaviors in only one category for each element. By **element**, we refer either to a unit of behavior, such as a sentence in a conversation, or to a time interval. To illustrate, suppose we were studying the problem-solving behavior of a group of public health workers discussing a new intervention for the homeless. Our category system involves eight categories: (1) seeks information, (2) gives information, (3) describes problem, (4) offers suggestion, (5) opposes suggestion, (6) supports suggestion, (7) summarizes, and (8) miscellaneous. Observers would be required to classify every group member's contribution—using, for example, each sentence as the element—in terms of one of these eight categories.

Another approach with exhaustive systems is to categorize relevant behaviors at regular time intervals. For example, in a category system for infants' motor activities, the researcher might use 10-second time intervals as the element; observers would categorize infant movements within 10-second periods.

Rating Scales

The major alternative to a checklist for recording structured observations is a **rating scale** that requires observers to rate a phenomenon along a descriptive continuum that is typically bipolar. The ratings are quantified for subsequent analysis.

Observers may be required to rate behaviors or events at specified intervals throughout the observational period (e.g., every 5 minutes). Alternatively, observers may rate entire events or transactions after observations are completed. Postobservation ratings require observers to integrate a number of activities and to judge which point on a scale most closely fits their interpretation of the situation. For example, suppose we were observing children's behavior during a scratch test for allergies. After each session, observers might be asked to rate the children's overall anxiety during the procedure on a graphic rating scale such as the following:

1	2	3	4	5	6	7
Extremely calm			Neither calm nor nervous			Extremely nervous

Rate how calm or nervous the child appeared to be during the procedure.

TIP: Global observational rating scales are sometimes included at the end of structured interviews. For example, in a study of the health problems of nearly 4,000 low-income mothers, interviewers were asked to observe and rate the safety of the home environment with regard to potential health hazards to the children on a five-point scale, from completely safe to extremely unsafe (Polit et al., 2001).

Rating scales can also be used as an extension of checklists, in which observers not only record the occurrence of a behavior, but also rate some qualitative aspect of it, such as its intensity. A good example is Weiss's (1992) Tactile Interaction Index (TII) for observing patterns of interpersonal touch. The TII comprises four dimensions: location (part of body touched, such as arm, abdomen), action (the specific gesture used, such as grabbing, hitting, patting); duration (temporal length of the touch), and intensity. Observers using the index must both classify the nature and duration of the touch *and* rate intensity on a four-point scale: light, moderate,

strong, and deep. When rating scales are coupled with a category scheme, considerable information about a phenomenon can be obtained, but it places an immense burden on observers, particularly if there is extensive activity.

Example of observational ratings: The NEECHAM Confusion Scale, an observational measure to detect the presence and severity of acute confusion, relies on ratings of behavior. For example, one rating concerns alertness/responsiveness, and the ratings are from 0 (responsiveness depressed) to 4 (full attentiveness). The NEECHAM has been used for both clinical and research purposes. For example, McCaffrey (2009) used NEECHAM scores to assess the effects of a music intervention on confusion in older adults after surgery.

TIP: It is usually useful to spend a period of time with participants before the actual observation and recording of data. Having a warm-up period helps to relax people (especially if audio or video equipment is being used) and can be helpful to observers (e.g., if participants have a linguistic style to which observers must adjust, such as a strong regional accent).

Constructing Versus Borrowing Structured Observational Instruments

As with self-report instruments, we encourage researchers to search for available observational instruments, rather than designing one themselves. The use of an existing instrument not only saves considerable work and time, but also facilitates comparisons among studies.

A few source books describe available observational instruments for certain research applications (e.g., Frank-Stromberg & Olsen, 2004), but the best source for such instruments is recent research literature on the study topic. For example, if you wanted to conduct an observational study of infant pain, a good place to begin would be recent research on this or similar topics to obtain information on how infant pain was operationalized.

Sampling for Structured Observations

Researchers must decide how, when, and for how long structured observational instruments will be used. Observations are usually done for a specific amount of time, and the amount of time is standardized across participants.

Sometimes *sampling* is needed so as to obtain representative examples of behaviors without having to observe for prolonged periods. Observational sampling concerns the selection of *behaviors* (or conversational segments) to be observed, not the selection of participants.

Time sampling involves the selection of time periods during which observations will occur. The time frames may be systematically selected (e.g., 60 seconds at 5-minute intervals) or selected at random. For example, suppose we were studying mothers' interactions with their children in a playground. During a 1-hour observation period, we sample moments to observe, rather than observing the entire session. Let us say that observations are made in 3-minute segments. If we used systematic sampling, we would observe for 3 minutes, then cease observing for a prespecified period, say 3 minutes. With this scheme, a total of ten 3-minute observations would be made. A second approach is to sample randomly 3-minute periods from the total of 20 such periods in an hour; a third is to use all 20 periods. Decisions about the length and number of periods for creating a good sample must be consistent with research aims. In establishing time units, a key consideration is determining a psychologically meaningful time frame. Pretesting and experimentation with different sampling plans is usually necessary.

Example of time sampling: Robb and colleagues (2008) tested the effect of active music engagement on stress and coping behaviors in children with cancer. Participating children received one of three interventions (active music engagement, music listening, or audio storybooks) and were then videotaped. Observers coded selected time segments (10 seconds, followed by 5-second segments) for facial affect, active engagement, and initiation.

Event sampling uses integral behavior sets or events for observation. Event sampling requires that the investigator either have knowledge about

the occurrence of events, or be in a position to wait for (or arrange) their occurrence. Examples of integral events suitable for event sampling include shift changes of hospital nurses or cast removals of pediatric patients. This approach is preferable to time sampling when events of interest are infrequent and are at risk of being missed. Still, when behaviors and events are frequent, time sampling has the virtue of enhancing the representativeness of observed behaviors.

Example of event sampling: Bryanton and colleagues (2009) explored whether mothers' perceptions of their childbirth experiences predicted early parenting behaviors. Parenting behaviors were observed during a feeding interaction when the infants were 1 month old.

Technical Aids in Observations

A wide array of technical devices is available for recording behaviors and events, making analysis or categorization at a later time possible. When the target behavior is auditory, recordings can be used to obtain a permanent record. Technological advances have vastly improved the quality, sensitivity, and unobtrusiveness of recording equipment. Auditory recordings can also be subjected to computerized speech software analysis to obtain objective quantitative measures of certain features of the recordings (e.g., volume, pitch).

Videotaping can be used when visual records are desired. In addition to being permanent, videotapes can capture complex behaviors that might elude on-the-spot observers. Visual records are also more capable than the naked eye of capturing fine units of behavior, such as micromomentary facial expressions. Videotapes make it possible to check the accuracy of coders and so are useful as a training aid. Finally, it is easier to conceal a camera than a human observer. Video records also have a few drawbacks, some of which are technical, such as lighting requirements, lens limitations, and so on. Sometimes the camera angle can present a lopsided view of an event. Also, some participants may be especially self-conscious in front of a video camera. Still, for many applications, permanent visual records offer unparalleled opportunities to expand the scope of observational studies. Haidet and colleagues (2009) offer valuable advice on improving data quality of video-recorded observations.

There is a growing technology for assisting with the encoding and recording of observations. For example, there is equipment that permits observers to enter observational data directly into a computer as the observation occurs, and in some cases, the equipment can record physiologic data concurrently.

Example of using equipment: Brown and colleagues (2009) developed and evaluated an observation system to assess mother–infant feeding interaction relevant to infant neuro-behavior regulation. In developing the system, videotapes of feeding sessions were digitized and stored on the computer so they could be opened for coding. They used a computer-based system (Observer) that offered a means of systematically observing and recording behavior as it occurred in real time. Coding was done by replaying the digitized video recording and entering observational codes into the computer.

Structured Observations by Nonresearch Observers

The observations discussed thus far are made and recorded by research team members. Sometimes, however, researchers ask people not connected with the research to provide structured data, based on their observations of the characteristics or behaviors of others. This method has much in common (in terms of format and scoring) with self-report scales; the primary difference is that the person completing the scale is asked to describe the attributes and behaviors of *another* person, based on observations of that person. For example, a mother might be asked to describe the behavior problems of her preschool child or staff nurses might be asked to evaluate the functional capacity of nursing home residents.

Obtaining observational data from nonresearchers is economical compared with using trained observers. For example, observers might have to watch children for hours or days to describe the nature and intensity of behavior problems, whereas parents or teachers could do this readily. Some behaviors might never lend themselves to outsider observation because of reactivity, occurrence in private situations, or infrequency (e.g., sleepwalking).

On the other hand, such methods may have the same problems as self-report scales (e.g., response-set bias) in addition to observer bias. Observer bias may in some cases be extreme, such as may happen when parents provide information about their children. Nonresearch observers are typically not trained, and interobserver agreement usually cannot be assessed. Thus, this approach has some problems but will continue to be used because, in many cases, there are no alternatives.

Example of observations by nonresearch personnel: Conrad and Altmaier (2009) studied the relationship between social support and levels of adjustment in children with cancer who attended a residential summer camp. Adjustment was measured by having parents complete the Child Behavior Checklist.

Evaluation of Structured Observation

Structured observation is an important data collection method, particularly for recording aspects of people's behaviors when they are not capable of describing them reliably in self-reports. Observational methods are particularly valuable for gathering data about infants and children, older people who are confused or agitated, or people whose communication skills are impaired.

Observations, like self-reports, are vulnerable to biases. One source of bias comes from those being observed. Participants may distort their behaviors in the direction of "looking good." They may also behave atypically because of their awareness of being observed, or their shyness in front of strangers or a camera.

Biases can also reflect human perceptual errors. Observation and interpretation are demanding tasks, requiring attention, perception, and conception. To accomplish these activities in a completely objective fashion is challenging and perhaps impossible. Biases are especially likely to operate when a high degree of observer inference is required.

Several types of observational bias are particularly common. One bias is the **enhancement of contrast effect**, in which observers distort observations in the direction of dividing content into clear-cut entities. The converse effect—a bias toward **central tendency**—occurs when extreme events are distorted toward a middle ground. With **assimilatory biases**, observers distort observations in the direction of identity with previous inputs. This bias would have the effect of miscategorizing information in the direction of regularity and orderliness. Assimilation to the observer's expectations and attitudes also occurs.

Rating scales are also susceptible to bias. The **halo effect** is the tendency of observers to be influenced by one characteristic in judging other, unrelated characteristics. For example, if we formed a positive general impression of a person, we might rate that person as intelligent, loyal, and dependable simply because these traits are positively valued. Ratings may reflect observers' personality. The **error of leniency** is the tendency for observers to rate everything positively, and the **error of severity** is the contrasting tendency to rate too harshly.

The careful construction and pretesting of checklists and rating scales, and the proper training and preparation of observers, play an important role in minimizing biases. To become a good instrument for collecting observational data, observers must be trained to observe in a manner that maximizes accuracy. Even when the lead researcher is the primary observer, self-training and dry runs are essential. The setting during the trial period should resemble as closely as possible the settings that will be the focus of actual observations.

Ideally, training should include practice sessions in which the comparability of observers' recordings

is assessed. That is, two or more independent observers should watch a trial situation, and observational coding should then be compared. Procedures for assessing the **interrater reliability** of structured observations are described in the next chapter.

TIP: Observations should be made in a neutral, nonjudgmental manner. People being observed are more likely to behave atypically if they think they are being critically appraised. Even positive cues (such as nodding approval) should be withheld because approval may induce repetition of a behavior that might not otherwise have occurred.

BIOPHYSIOLOGIC MEASURES

Settings in which nurses work are typically filled with a wide variety of technical instruments for measuring physiologic functions. It is beyond the scope of this book to describe the many kinds of biophysiologic measures available to nurse researchers. Our goals are to present an overview of biophysiologic measures, to illustrate their use in research, and to note considerations in decisions to use them.

Purposes of Collecting Biophysiologic Data

Clinical nursing studies involve biophysiologic instruments both for creating independent variables (e.g., a biofeedback intervention) and for measuring outcomes. For the most part, our discussion focuses on the use of biophysiologic measures as dependent (outcome) variables. Examples of the purposes of collecting biophysiologic data include the following:

1. *Studies of basic biophysiologic processes that have relevance for nursing care*. These studies involve healthy participants or an animal species. For example, Dorsey and colleagues (2009) studied mechanisms underlying painful peripheral neuropathy in the treatment of HIV using a whole-genome microassay screen with a mouse model.
2. *Descriptions of the physiologic consequences of nursing and healthcare.* These studies do not focus on specific interventions, but rather are designed to learn how standard procedures affect patients' physiologic outcomes. For example, Kang and colleagues (2009) tracked immune recovery (e.g., natural killer cell activity) in the 12 months following cancer treatment among women with early-stage breast cancer.
3. *Evaluations of a specific nursing intervention*. Some studies involve testing the effects of a new intervention, usually in comparison with standard methods of care or alternative interventions. Typically, these studies test the hypothesis that the innovation will result in improved biophysiologic outcomes among patients. As an example, Yeo (2009) tested the effects of a walking versus stretching exercise on preeclampsia risk factors such as heart rate and blood pressure in sedentary pregnant women.
4. *Assessments of products or clinical procedures*. Some studies evaluate products designed to enhance patient health or comfort, or test alternative products and procedures. For example, Mathew and colleagues (2009) collected central catheter blood samples using three alternative methods and compared blood culture results.
5. *Studies of the correlates of physiologic functioning in patients with health problems*. Researchers study possible antecedents and consequences of biophysiologic outcomes to gain insight into potential treatments or modes of care. Nurse researchers have also studied biophysiologic outcomes in relation to social or psychological characteristics. As an example, Neira and colleagues (2009) studied the association between glucose metabolism and cardiometabolic risk factors in Hispanics at risk for metabolic syndrome.

Types of Biophysiologic Measures

Physiologic measurements are either in vivo or in vitro. **In vivo measurements** are performed directly in or on living organisms. Examples include measures of oxygen saturation, blood pressure, and body temperature. An **in vitro measurement**, by contrast, is performed outside the organism's body, as in the case of measuring serum potassium concentration in the blood.

In vivo measures often involve the use of highly complex instrumentation systems, involving (for example) a stimulus, sensing equipment (e.g., transducers), signal-conditioning equipment to reduce interference, display equipment, and recording and data processing equipment. In vivo instruments have been developed to measure all bodily functions, and technological improvements continue to advance our ability to measure biophysiologic phenomena more accurately, more conveniently, and more rapidly than ever before. The uses to which such instruments have been put by nurse researchers are richly diverse.

Example of a study with in vivo measures: Ayhan and colleagues (2009) randomly assigned patients undergoing a thyroidectomy to two oxygen-delivery methods (face mask and nasal cannula) and then assessed the effect on peripheral oxygen saturation, measured by pulse oximetry every 5 minutes for 30 minutes.

With in vitro measures, data are gathered by extracting physiologic material from people and submitting it for laboratory analysis. Nurse researchers may or may not be involved in the extraction of the material; however, the analysis is normally done by specialized laboratory technicians. Usually, each laboratory establishes a range of normal values for each measurement, and this information is critical for interpreting the results. Several classes of laboratory analysis have been used by nurse researchers, including chemical measurements (e.g., measures of potassium levels), microbiologic measures (e.g., bacterial counts), and cytologic *or* histologic measures (e.g., tissue biopsies). Laboratory analyses of blood and urine samples are the most frequently used in vitro measures in nursing investigations.

Example of a study with in vitro measures: Choi and Rankin (2009) studied factors influencing glucose control in Korean immigrants with type 2 diabetes. A finger stick blood test was used to assess levels of glycosylated hemoglobin (HbA1c).

Selecting a Biophysiologic Measure

The most basic issue in selecting a physiologic measure is whether it will yield good information about research variables. In some cases, researchers need to consider whether the variable should be measured by observation or self-report instead of (or in addition to) using biophysiologic equipment. For example, stress could be measured by asking people questions (e.g., using the State–Trait Anxiety Inventory), by observing their behavior during exposure to stressful stimuli, or by measuring heart rate, blood pressure, or levels of adrenocorticotropic hormone in urine samples.

Several other considerations should be kept in mind in selecting a biophysiologic measure. Some key questions include the following:

- Is the equipment or laboratory analysis you need readily available to you? If not, can it be borrowed, rented, or purchased?
- Can you operate the required equipment and interpret its results, or do you need training? Are resources available to help you with operation and interpretation?
- Will you have difficulty obtaining permission to use the equipment from an Institutional Review Board or other institutional authority?
- Do your activities during data collection permit you to record data simultaneously, or do you need an instrument system with recording equipment (or a research assistant)?
- Is a single measure of the dependent variable sufficient, or are multiple measures needed for a reliable estimate? If the latter, what burden does this place on participants?
- Are your measures likely to be influenced by reactivity (i.e., participants' awareness of their status)? If so, can alternative or supplementary nonreactive measures be identified, or can the extent of reactivity bias be assessed?

- Is the measure you plan to use sufficiently accurate and sensitive to variation?
- Are you thoroughly familiar with rules and safety precautions, such as grounding procedures, especially when using electrical equipment?

Evaluation of Biophysiologic Measures

Biophysiologic measures offer the following advantages to nurse researchers:

- Biophysiologic measures are accurate and precise compared with psychological measures (e.g., self-report measures of anxiety).
- Biophysiologic measures are objective. Two nurses reading from the same sphygmomanometer are likely to obtain the same blood pressure measurements, and two different sphygmomanometers are likely to produce identical readouts. Patients cannot easily distort measurements of biophysiologic functioning deliberately.
- Biophysiologic instruments provide valid measures of targeted variables: thermometers can be depended on to measure temperature and not blood volume, and so forth. For self-report and observational measures, it is often more difficult to be certain that the instrument is really measuring the target concept.

Biophysiologic measures also have a few disadvantages:

- The cost of collecting some types of biophysiologic data may be low or nonexistent, but when laboratory tests are involved, they may be more expensive than other methods (e.g., assessing smoking status by means of cotinine assays versus self-report).
- The measuring tool may affect the variables it is attempting to measure. The presence of a sensing device, such as a transducer, located in a blood vessel partially blocks that vessel and, hence, alters the pressure–flow characteristics being measured.
- Energy must often be applied to the organism when taking the biophysiologic measurements; extreme caution must continually be exercised to avoid the risk of damaging cells by high-energy concentrations.

The difficulty in choosing biophysiologic measures for nursing studies lies not in their shortage, nor in their questionable utility, nor in their inferiority to other methods. Indeed, they are plentiful, often highly reliable and valid, and extremely useful in clinical nursing studies. Care must be exercised, however, in selecting instruments or laboratory analyses with regard to practical, ethical, medical, and technical considerations.

IMPLEMENTING A DATA COLLECTION PLAN

Data quality in a quantitative study is affected by both the data collection plan and how the plan is implemented.

Selecting Research Personnel

An important decision concerns who will actually collect the research data. In small studies, the lead researcher usually collects the data personally. In larger studies, however, this may not be feasible. When data are collected by others, it is important to select appropriate people. In general, they should be neutral agents through whom data passes—that is, their characteristics or behavior should not affect the substance of the data. Some considerations that should be kept in mind when selecting research personnel are as follows:

- *Experience*. Research staff ideally have had prior experience collecting data (e.g., prior interviewing experience). If this is not feasible, look for people who can readily acquire the necessary skills (e.g., an interviewer should have good verbal and social skills).
- *Congruity with sample characteristics*. If possible, data collectors should match participants with respect to racial or cultural background and gender. The greater the sensitivity of the questions, the greater the desirability of matching characteristics.

- *Unremarkable appearance*. Extremes of appearance should be avoided. For example, data collectors should not dress very casually (e.g., in shorts and tee shirts), nor formally (e.g., in designer clothes). Data collectors should not wear anything that conveys their political, social, or religious views.
- *Personality*. Data collectors should be pleasant (but not effusive), sociable (but not overly talkative), and nonjudgmental (but not unfeeling about participants' lives). The goal is to have nonthreatening data collectors who can put participants at ease.

In some situations, researchers cannot select research personnel. For example, the data collectors may be staff nurses employed at a hospital. Training of the data collection staff is particularly important in such situations. Even if there are no additional data collection staff, researchers should self-monitor their demeanor and prepare for their role with care.

Training Data Collectors

Depending on prior experience, training will need to cover both general procedures (e.g., how to probe in an interview) and ones specific to the study (e.g., how to ask a particular question). Training can often be done in a single day, but complex projects require more time. The lead researcher is usually the best person to conduct the training and to develop training materials.

Data collection protocols usually are a good foundation for a **training manual**. The manual normally includes background materials (e.g., the study aims), general instructions, specific instructions, and copies of all data forms.

TIP: A table of contents for a training manual is included in the Toolkit of the accompanying *Resource Manual.* Models for some of the sections in this table of contents (a section on avoiding interviewer bias and another on how to probe) are also available in the Toolkit. If you are collecting the data yourself, you may not need a training manual, but you should learn techniques of professional interviewing.

The agenda for the training should cover the content of the training manual, elaborating on any portion that is especially complex. Training usually includes demonstrations of fictitious data collection sessions, performed either live or on videotape. Finally, training usually involves having trainees do trial runs of data collection (e.g., *mock interviews*) in front of the trainers to demonstrate their understanding of the instructions. Thompson and colleagues (2005) provide some additional tips relating to the training of research personnel.

Example of data collector training: In a two-wave panel study of the health of nearly 4,000 low-income families, Polit and colleagues (2001) trained about 100 interviewers in 4 research sites. Each training session lasted 3 days, including a half day of training on the use of CAPI. At the end of the training, several trainees were not kept on as interviewers because they were not skillful in mastering their assignments.

CRITIQUING STRUCTURED METHODS OF DATA COLLECTION

The goal of a data collection plan is to produce data that are of exceptional quality. Every decision researchers make about data collection methods and procedures is likely to affect data quality, and hence overall study quality. These decisions should be critiqued in evaluating the study's evidence to the extent possible. The critiquing guidelines in Box 13.3 focus on global decisions about the design and implementation of a data collection plan. Unfortunately, data collection procedures are often not described in detail in research reports, owing to space constraints in journals. A full critique of data collection plans is rarely feasible.

A second set of critiquing guidelines is presented in Box 13.4. These questions focus on the specific methods of collecting research data in quantitative studies. Further guidance on drawing conclusions about data quality in quantitative studies is provided in the next chapter.

BOX 13.3 Guidelines for Critiquing Data Collection Plans in Quantitative Studies

1. Was the collection of data using structured methods (in contrast with unstructured methods) consistent with study aims?
2. Were the right methods used to collect the data (self-report, observation, etc.)? Was triangulation of methods used appropriately? Should supplementary data collection methods have been used to enrich the data available for analysis?
3. Was the right amount of data collected? Were data collected to address the varied needs of the study? Was *too much* data collected in terms of burdening study participants—and, if so, how might this have affected data quality?
4. Did the researcher select good instruments, in terms of congruence with underlying constructs, data quality, reputation, efficiency, and so on? Were new instruments developed without a justifiable rationale?
5. Were data collection instruments adequately pretested?
6. Did the report provide sufficient information about data collection procedures?
7. Who collected the data? Were data collectors judiciously chosen, with traits that were likely to enhance data quality?
8. Was the training of data collectors described? Was the training adequate? Were steps taken to improve data collectors' ability to elicit or produce high-quality data, or to monitor their performance?
9. Where and under what circumstances were data gathered? Was the setting for data collection appropriate?
10. Were other people present during data collection? Could the presence of others have resulted in any biases?
11. Were data collectors blinded to study hypotheses or to participants' group status?

BOX 13.4 Guidelines for Critiquing Structured Data Collection Methods

1. If self-report methods were used, did the researcher make good decisions about the specific method used to solicit self-report information (e.g., mix of open- and closed-ended questions, use of composite scales, and so on)?
2. Was the instrument package adequately described in terms of reading level of the questions, length of time to complete it, modules included, and so on?
3. Was the mode of obtaining the self-report data appropriate (e.g., in-person interviews, mailed SAQs, Internet questionnaires, and so on)?
4. Were self-report data gathered in a manner that promoted high-quality and unbiased responses (e.g., in terms of privacy, efforts to put respondents at ease, and so on)?
5. If observational methods were used, did the report adequately describe the specific constructs that were observed? What was the unit of observation, and was this appropriate?
6. Was a category system or rating system used to organize and record observations? Was the category system exhaustive? How much inference was required of the observers? Were decisions about exhaustiveness and degree of observer inference appropriate?
7. What methods were used to sample observational units? Was the sampling approach a good one, and did it likely yield a representative sample of behavior?
8. To what degree were observer biases controlled or minimized?
9. Were biophysiologic measures used in the study, and was this appropriate? Did the researcher appear to have the skills necessary for proper interpretation of biophysiologic measures?

RESEARCH EXAMPLE

In the study described next, a variety of data collection approaches was used to measure study variables.

Study: Predicting children's response to distraction from pain (Dr. Ann McCarthy & Dr. Charmaine Kleiber, Principal Investigators, NINR grant 1-R01-NR005269).

Statement of Purpose: Drs. McCarthy and Kleiber developed and tested an intervention to train parents as coaches to distract their children during insertion of an intravenous (IV) catheter. The overall study purpose was to test the effectiveness of the intervention in reducing children's pain and distress, to identify factors that predicted which children benefited from the distraction, and to identify characteristics of parents who were successful in distracting their children.

Design: In this multisite clinical trial, 542 parents were randomly assigned to an intervention group or a usual-care control group. Their children, aged 4 to 10, were scheduled to undergo an IV insertion for a diagnostic medical procedure. Parents in the intervention group received 15 minutes of training regarding effective methods of distraction before the child's IV insertion.

Data Collection Plan: The researchers collected a wide range of data both prior to and following the intervention and IV procedure, using self-report, observational, and biophysiologic measures. Their data collection plan included the use of formal instruments for describing sample characteristics, for assessing key outcomes of children's pain and distress, for measuring parent and child factors they hypothesized would predict the intervention's effectiveness, for capturing characteristics of the IV procedure, and for evaluating treatment fidelity in terms of parental success with distraction coaching. The researchers undertook a thorough literature review to identify factors influencing children's responses to a painful procedure, and developed a model that guided their data collection efforts. Before undertaking the full-scale study, the instruments were pilot tested (Kleiber & McCarthy, 2006). The pilot test was used to assess whether the instruments were understandable, to evaluate the quality of data they would yield, and to explore interrelationships among study variables. The researchers noted "the value of evaluating instruments prior to the initiation of a larger study" (p. 104). Because of the extensiveness of their data collection plan, we describe only a few specific measures here.

Self-Report Instruments: Both parents and children provided self-report data. For example, scores on the Oucher Scale, a self-report measure of children's pain, were used as an outcome variable. Children also reported their level of anxiety on a visual analog scale. Another child self-report instrument (Child Behavioral Style Scale) measured their coping style, using a vignette-type approach with four stressful scenarios. Parents completed self-administered questionnaires that incorporated scales to measure parenting style (Parenting Dimensions Inventory) and anxiety (State-Trait Anxiety Inventory). They also completed instruments that described their children's temperament (Dimensions of Temperament Survey).

Observational Instruments: A research assistant videotaped the parent and the child during the time they were in the treatment room. Videotapes were entered into a computerized video editing program and divided into 10-second intervals for analysis. The authors coded the parents' behavior in terms of the quality and frequency of distraction coaching, using an observational instrument that the researchers carefully developed, the Distraction Coaching Index (Kleiber et al., 2007). The videotapes were also used to code the children's behavioral distress, using the Observation Scale of Behavioral Distress.

Biophysiologic Measures: Children's stress was also measured using salivary cortisol levels. The chew-and-spit technique was used to collect salivary samples. Children chewed a piece of sugarless gum as a salivary stimulant. After discarding the gum, the children spat saliva into a collection tube. Each child provided four salivary cortisol samples: before IV insertion, 20 minutes after IV insertion, and two home samples to assess the child's baseline cortisol levels. Care was taken to ensure the integrity of the samples and to control conditions under which they were obtained (McCarthy et al., 2009).

Key Findings: Results from this extensive study are just appearing in the literature. Early published results have indicated that parents in the intervention group had significantly higher scores than those in the control group for distraction coaching frequency and quality (Kleiber et al., 2007). The researchers also found, using data from control group children, that baseline cortisol levels were lower than levels obtained in the clinics, and that cortisol levels increased following IV insertion, supporting the utility of cortisol levels as a measure of stress response (McCarthy et al., 2009).

SUMMARY POINTS

- Quantitative researchers typically develop a detailed **data collection plan** before they begin to collect their data. For structured data, researchers use formal data collection **instruments** that place constraints on those collecting data and those providing them.
- An early step in developing a data collection plan is the identification and prioritization of data needs. After data needs have been identified, measures of the variables must be located. The selection of existing instruments should be based on such considerations as conceptual suitability, data quality, cost, population appropriateness, and reputation.
- Even when existing instruments are used, the instrument package should be **pretested** to assess its length, clarity, and overall adequacy.
- Structured self-report instruments (**interview schedules** or **questionnaires**) may include open- or closed-ended questions. **Open-ended questions** permit respondents to reply in narrative fashion, whereas **closed-ended** (or **fixed-alternative**) **questions** offer *response alternatives* from which respondents must choose.
- Types of closed-ended questions include (1) **dichotomous questions**, which require a choice between two options (e.g., yes/no); (2) **multiple-choice questions**, which offer a range of alternatives; (3) **rank-order questions**, in which respondents are asked to rank concepts on a continuum; (4) **forced-choice questions**, which require respondents to choose between two competing positions; (5) **rating questions**, which ask respondents to make judgments along a bipolar dimension; (6) **checklists** that have several questions with the same response format; and (7) **visual analog scales** (VASs), which are used to measure subjective experiences such as pain. **Event history calendars** and diaries are used to capture data about the occurrence of events.
- Composite psychosocial **scales** are multiple-item self-report tools for measuring the degree to which individuals possess or are characterized by target attributes.
- **Likert scales** (**summated rating scales**) comprise a series of statements (**items**) about a phenomenon. Respondents typically indicate degree of agreement or disagreement with each statement; a total score is computed by summing item scores, each of which is scored for the intensity and direction of favorability expressed.
- **Semantic differentials** (SDs) consist of a series of bipolar rating scales on which respondents indicate reactions toward a phenomenon; scales can measure an evaluative (e.g., good/bad), activity (e.g., active/passive), or potency (e.g., strong/weak) dimension.
- **Q sorts**, in which people sort a set of card statements into piles according to specified criteria, can be used to measure attitudes, personality, and other psychological traits.
- **Vignettes** are brief descriptions of an event or situation to which respondents are asked to react. They are used to assess respondents' perceptions, hypothetical behaviors, or decisions.
- Questionnaires are less costly and time-consuming than interviews, offer the possibility of anonymity, and run no risk of interviewer bias. Interviews have higher response rates, are suitable for a wider variety of people, and yield richer data than questionnaires.
- Data quality in interviews depends on interviewers' interpersonal skills. Interviewers must put respondents at ease and build rapport, and need to be skillful at *probing* for additional information when respondents give incomplete responses.
- Group administration is the most economical way to distribute questionnaires. Another approach is to mail them, but this method tends to have low **response rates**, which can result in bias. Questionnaires can be distributed via the Internet, most often as a **web-based survey** that is accessed through a hypertext link. Several techniques, such as **follow-up reminders** and good **cover letters**, increase response rates to questionnaires.
- Structured self-reports are vulnerable to the risk of reporting biases. **Response set biases** reflect the tendency of some people to respond to

questions in characteristic ways, independently of content. Common response sets include **social desirability, extreme response,** and **acquiescence (yea-saying).**

- Structured observational methods impose constraints on observers, to enhance the accuracy and objectivity of observations and to obtain an adequate representation of phenomena of interest.
- **Checklists** are used in observations to recording the occurrence or frequency of designated behaviors, events, or characteristics. Checklists are based on **category systems** for encoding observed phenomena into discrete categories.
- With **rating scales**, observers rate phenomena along a dimension that is typically bipolar (e.g., passive/aggressive); ratings are made either at specific intervals (e.g., every 5 minutes) or after observations are completed.
- **Time sampling** involves the specification of the duration and frequency of observational periods and intersession intervals. **Event sampling** selects integral behaviors or events of a special type for observation.
- Observational methods are an excellent way to operationalize some constructs, but are subject to various biases. The greater the degree of observer inference, the more likely that distortions will occur. The most prevalent observer biases include the **enhancement of contrast effect, central tendency bias**, the **halo effect, assimilatory biases, errors of leniency**, and **errors of severity**.
- **Biophysiologic measures** comprise **in vivo measurements** (those performed within or on living organisms, like blood pressure measurement) and **in vitro measurements** (those performed outside the organism's body, such as blood tests).
- Biophysiologic measures are objective, accurate, and precise, but care must be taken in using such measures with regard to practical, technical, and ethical considerations.
- When researchers cannot collect the data without assistance, they should carefully select data collection staff and formally train them.

STUDY ACTIVITIES

Chapter 13 of the *Resource Manual for Nursing Research: Generating and Assessing Evidence for Nursing Practice, 9th edition*, offers exercises and study suggestions for reinforcing concepts presented in this chapter. In addition, the following study questions can be addressed:

1. Suppose you were planning to conduct a statewide study of the work plans and intentions of nonemployed registered nurses in your state. Would you ask mostly open-ended or closed-ended questions? Would you adopt an interview or questionnaire approach? If a questionnaire, how would you distribute it?
2. Suppose that the study of nonemployed nurses were done by a mailed questionnaire. Draft a cover letter to accompany it.
3. A nurse researcher is planning to study temper tantrums displayed by hospitalized children. Would you recommend using a time sampling approach? Why or why not?

STUDIES CITED IN CHAPTER 13

Akhar-Danesh, N., Baxter, P., Valaitis, R., Stanyon, W., & Sproud, S. (2009). Nurse faculty perceptions of simulation use in nursing education. *Western Journal of Nursing Research, 31*, 312–329.

Alpert, P., Miller, S., Wallmann, H., Havey, R., Cross, C., Chevalia, T., Gillis, C. B., & Kodandapari, K. (2009). The effect of modified jazz dance on balance, cognition, and mood in older adults. *Journal of the American Academy of Nurse Practitioners, 21*, 108–115.

Ayhan, H., Iyigun, E., Tastan, S., Orhan, M., & Ozturk, E. (2009). Comparison of two different oxygen delivery methods in the early postoperative period. *Journal of Advanced Nursing, 65*, 1237–1247.

Berger, A., Treat Marunda, H., & Agrawal, S. (2009). Influence of menopausal status on sleep and hot flashes throughout breast adjuvant chemotherapy. *Journal of Obstetric, Gynecologic, & Neonatal Nursing, 38*, 353–366.

Brown, L., Thoyre, S., Pridham, K., & Schubert, C. (2009). The Mother-Infant Feeding Tool. *Journal of Obstetric, Gynecologic, & Neonatal Nursing, 38*, 491–503.

Bryanton, J., Gagnon, A., Hatem, M., & Johnston, C. (2009). Does perception of the childbirth experience predict women's early parenting behaviors? *Research in Nursing & Health, 32*, 191–203.

Choi, S., & Rankin, S. (2009). Glucose control in Korean immigrants with type 2 diabetes. *Western Journal of Nursing Research, 31*, 347–363.

Choi, J. Y., & Hwang, S. Y. (2009). Factors associated with health-related quality of life among low-compliance asthmatic adults in Korea. *Research in Nursing & Health, 32*, 140–147.

Conrad, A., & Altmaier, E. (2009). Specialized summer camp for children with cancer: Social support and adjustment. *Journal of Pediatric Oncology Nursing, 26*, 150–157.

Dirksen, S. R., Epstein, D., & Hoyt, M. (2009). Insomnia, depression, and distress among outpatients with prostate cancer. *Applied Nursing Research, 22*, 154–158.

Dorsey, S., Leitch, C., Renn, C., Lessans, S., Smith, B., Wang, X., & Dionne, R. (2009). Genome-wide screen identifies drug-induced regulation of the gene giant axonal neuropathy in a mouse model of antiretroviral-induced painful peripheral neuropathy. *Biological Research for Nursing, 11*, 7–16.

Foreman, S. W., Thomas, K., & Blackburn, S. (2008). Individual and gender differences matter in preterm infant state development. *Journal of Obstetric, Gynecologic, & Neonatal Nursing, 37*, 657–665.

Griffin, R., Polit, D., & Byrne, M. (2007). The effects of children's gender, race, and attractiveness on nurses' pain management decisions. *Research in Nursing & Health, 30*, 655–666.

Gross-King, M., Booth-Jones, M., & Couluris, M. (2008). Neurocognitive impairment in children treated for cancer. *Journal of Pediatric Oncology Nursing, 25*, 227–232.

Johnston, C., Filion, F., Campbell-Yeo, M., Goulet, C., Bell, L., McNaughton, K., Byron, J., Aita, M., Finley, G. A., Walker, C. D. (2008). Kangaroo mother care diminishes pain from heel lance in very preterm neonates. *BMC Pediatrics, 8*, 13.

Kang, D., Weaver, M., Park, N., Smith, B., McArdle, T., & Carpenter, J. (2009). Significant impairment in immune recovery after cancer treatment. *Nursing Research, 58*, 105–114.

Kleiber, C., & McCarthy, A. M. (2006). Evaluating instruments for a study on children's responses to a painful procedure when parents are distraction coaches. *Journal of Pediatric Nursing, 21*, 99–107.

Kleiber, C., McCathy, A. M., Hanrahan, K., Myers, L., & Weathers, N. (2007). Development of the Distraction Coaching Index. *Children's Healthcare, 36*, 219–235.

Kupferer, E., Dormire, S., & Becker, H. (2009). Complementary and alternative medicine use for vasomotor symptoms among women who have discontinued hormone therapy. *Journal of Obstetric, Gynecologic, & Neonatal Nursing, 38*, 50–59.

Liaw, J., Yang, L., Yuh, Y., & Yin, T. (2006). Effects of tub bathing procedures on preterm infants' behavior. *Journal of Nursing Research, 14*, 297–305.

Lynn, M., Morgan, J., & Moore, K. (2009). Development and testing of the Satisfaction in Nursing Scale. *Nursing Research, 58*, 166–174.

Mathew, A., Gasline, T., Dunning, K., & Ying, J. (2009). Central catheter blood sampling: the impact of changing the needleless caps prior to collection. *Journal of Infusion Nursing, 32*, 212–218.

McCaffrey, R. (2009). The effect of music on acute confusion in older adults after hip or knee surgery. *Applied Nursing Research, 22*, 107–112.

McCarthy, A. M., Hanrahan, K., Kleiber, C., Zimmerman, M. B., Lutgendorf, S., & Tsalikian, E. (2009). Normative salivary cortisol values and responsivity in children. *Applied Nursing Research, 22*, 54–62.

Neira, C., Hartig, M., Cowan, P., & Velasquez-Mieyer, P. (2009). The prevalence of impaired glucose metabolism in Hispanics with two or more risk factors for metabolic syndrome in the primary care setting. *Journal of the American Academy of Nurse Practitioners, 21*, 173–178.

Nyamathi, A., Berg, J., Jones, T., & Leake, B. (2005). Predictors of perceived health status of tuberculosis-infected homeless. *Western Journal of Nursing Research, 27*, 896–910.

Polit, D. F., London, A. S., & Martinez, J. M. (2001). *The health of poor urban women*. New York: MDRC.

Rempusheski, V. F., & O'Hara, C. (2005). Psychometric properties of the Granparent Perceptions of Family Scale (GPFS). *Nursing Research, 54*, 419–427.

Robb, S., Clair, A., Watanabe, M., Monahan, P., Azzouz, F., Stouffer, J., Ebberts, A., Darsie, E., Whitmer, C., Walker, J., Nelson, K., Hanson-Abromeit, D., Lane, D., & Hannan, A. (2008). A non-randomized controlled trial of the active music engagement (AME) intervention on children with cancer. *Psycho-oncology*, 17, 699–708.

Sarna, L., Bialous, S., Wells, M., Kotlerman, J., Wewers, M., & Froelicher, E. (2009). Frequency of nurses' smoking cessation interventions: Report from a national survey. *Journal of Clinical Nursing, 18*, 2066–2077.

Uitterhoeve, R., de Leeuw, J., Bensing, J., Heaven, C., Borm, G., deMulder, P., &van Achterberg, T. (2008). Cue-responding behaviours of oncology nurses in video-simulated interviews. *Journal of Advanced Nursing, 61*, 71–80.

Weiss, S. J. (1992). Measurement of the sensory qualities in tactile interaction. *Nursing Research, 41*, 82–86.

Yeo, S. (2009). Adherence to walking or stretching, and risk of preeclampsia in sedentary pregnant women. *Research in Nursing & Health, 32*, 379–390.

Methodologic and nonresearch references cited in this chapter can be found in a separate section at the end of the book.

14 Measurement and Data Quality

An ideal data collection procedure is one that captures a construct in a way that is accurate, truthful, and sensitive. Biophysiologic methods have a higher chance of success in attaining these goals than self-report or observational methods, but no method is flawless. In this chapter, we discuss criteria for evaluating the quality of data obtained with structured instruments.

We begin by discussing principles of measurement. Our discussion is based primarily on **classical measurement theory (CMT)**, the leading theory with regard to the measurement of affective constructs (i.e., constructs such as self-esteem or depression). An alternative measurement theory (*item response theory* or *IRT*) has gained in popularity, especially for measuring cognitive constructs (e.g., knowledge). We discuss IRT briefly in Chapter 15.

MEASUREMENT

Quantitative studies derive data through the measurement of variables. **Measurement** involves assigning numbers to represent the amount of an attribute present in an object or person, using a specified set of rules. Quantification and measurement go hand in hand. Attributes are not constant; they vary from day to day or from one person to another. Variability is presumed to be capable of a numeric expression signifying *how much* of an attribute is present. The purpose of assigning numbers is to differentiate between people with varying degrees of the attribute.

Rules and Measurement

Measurement involves assigning numbers according to rules. Rules for measuring temperature, weight, and other physical attributes are familiar to us. Rules for measuring many variables for nursing studies, however, have to be invented. Whether the data are collected by observation, self-report, or some other method, researchers must specify criteria for assigning numeric values to the characteristic of interest.

As an example, suppose we were studying parental attitudes toward dispensing condoms in school clinics, and we asked parents their extent of agreement with the following statement:

Teenagers should have access to contraceptives in school clinics.

- ❏ Strongly disagree
- ❏ Disagree
- ❏ Slightly disagree
- ❏ Neither agree nor disagree
- ❏ Slightly agree
- ❏ Agree
- ❏ Strongly agree

Responses to this question can be quantified by developing a system for assigning numbers to them. Note that *any* rule would satisfy the definition of measurement. We could assign the value of 30 to "strongly agree," 28 to "agree," 20 to "slightly agree," and so on, but there is no justification for doing so. In measuring attributes, researchers strive to use good, meaningful rules. Without *a priori* knowledge of the "distance" between response options, the most practical approach is to assign a 7 to "strongly agree" and a 1 to "strongly disagree." This rule would quantitatively differentiate, in increments of one point, among people with seven different opinions. Researchers seldom know in advance if their rules are the best possible. New measurement rules reflect hypotheses about how attributes vary. The adequacy of the hypotheses—that is, the worth of the instruments—needs to be assessed empirically.

Researchers try to link numeric values to reality. To state this goal more technically, measurement procedures are ideally isomorphic to reality. The term *isomorphism* signifies equivalence or similarity between two phenomena. An instrument cannot be useful unless the measurements resulting from it correspond with the real world.

To illustrate the concept of isomorphism, suppose a standardized test was administered to 10 students, who obtained the following scores: 345, 395, 430, 435, 490, 505, 550, 570, 620, and 640. These values are shown at the top of Figure 14.1. Suppose that in reality the students' true scores on a hypothetically perfect test were as follows: 360, 375, 430, 465, 470, 500, 550, 610, 590, and 670, shown at the bottom of Figure 14.1. Although not perfect, the test came close to representing true scores; only two people (H and I) were improperly ordered. This example illustrates a measure whose isomorphism with reality is high but improvable.

Researchers work with fallible measures. Instruments that measure psychosocial phenomena are less likely to correspond to reality than physical measures, but few instruments are error free.

Advantages of Measurement

What exactly does measurement accomplish? Consider how handicapped healthcare professionals would be in the absence of measurement. What would happen, for example, if there were no measures of blood pressure or temperature? Subjective evaluations of clinical outcomes would have to be used. A principal strength of measurement is that it removes subjectivity and guesswork. Because measurement is based on explicit rules, resulting information tends to be objective—that is, it can be independently verified. Two people measuring the weight of a person using the same scale would likely get identical results. Most measures incorporate mechanisms for minimizing subjectivity.

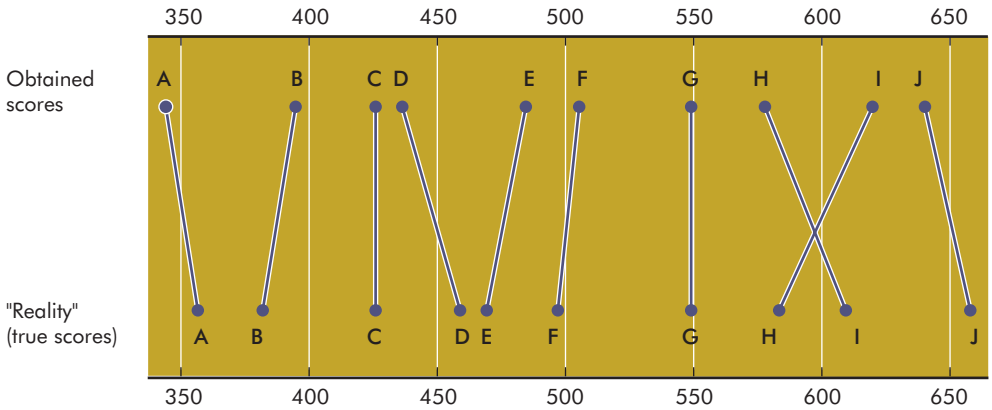


FIGURE 14.1 Relationship between obtained and true scores for a hypothetical set of test scores.

Measurement also makes it possible to obtain reasonably precise information. Instead of describing Nathan as "rather tall," we can depict him as being 6 feet 3 inches tall. With precise measures, researchers can differentiate among people with different degrees of an attribute.

Finally, measurement is a language of communication. Numbers are less vague than words and can communicate information more accurately. If a researcher reported that the average oral temperature of a sample of patients was "somewhat high," different readers might make different inferences about the sample's physiologic state. However, if the researcher reported an average temperature of 99.6°F, there would be no ambiguity.

Errors of Measurement

Procedures for obtaining measurements, as well as the objects being measured, are susceptible to influences that can alter the resulting data. Some influences can be controlled to a certain degree, and attempts should be made to do so, but such efforts are rarely completely successful.

Instruments that are not perfectly accurate yield measurements containing some error. Within classical measurement theory, an **observed** (or **obtained**) **score** can be conceptualized as having two parts—an error component and a true component. This can be written symbolically as follows:

$$\text{Obtained score} = \text{True score} \pm \text{Error}$$

or

$$X_O = X_T \pm X_E$$

The first term in the equation is an observed score—for example, a score on an anxiety scale. X_T is the value that would be obtained with an infallible measure. The **true score** is hypothetical—it can never be known because measures are *not* infallible. The final term is the **error of measurement**. The difference between true and obtained scores is the result of factors that distort the measurement.

Decomposing obtained scores in this manner highlights an important point. When researchers measure an attribute, they are also *measuring* attributes that are not of interest. The true score component is what they hope to isolate; the error component is a composite of other factors that are also being measured, contrary to their wishes. This concept can be illustrated with an exaggerated example. Suppose a researcher measured the weight of 10 people on a spring scale. As participants step on the scale, the researcher places a hand on their shoulders and applies pressure. The resulting measures (the X_Os) will be biased upward because scores reflect both actual weight (X_T) and pressure (X_E). Errors of measurement are problematic because their value is unknown and also because they often are variable. In this example, the amount of pressure applied likely would vary from one person to the next. In other words, the proportion of true score component in an obtained score varies from one person to the next.

Many factors contribute to errors of measurement. Some errors are random while others are systematic, reflecting *bias*. Common influences on measurement error include the following:

1. *Situational contaminants*. Scores can be affected by the conditions under which they are produced. A participant's awareness of an observer's presence (reactivity) is one source of bias. Environmental factors, such as temperature, lighting, and time of day, are potential sources of measurement error.
2. *Transitory personal factors*. A person's score can be influenced by such personal states as fatigue or mood. In some cases, such factors directly affect the measurement, as when anxiety affects pulse rate measurement. In other cases, personal factors alter scores by influencing people's motivation to cooperate, act naturally, or do their best.
3. *Response-set biases*. Relatively enduring characteristics of people can interfere with accurate measurements. Response sets such as social desirability or acquiescence are potential biases in self-report measures, particularly in psychological scales (Chapter 13).
4. *Administration variations*. Alterations in the methods of collecting data from one person to the next can result in score variations unrelated

to variations in the target attribute. For example, if some physiologic measures are taken before a feeding and others are taken after a feeding, then measurement errors can potentially occur.

5. *Instrument clarity*. If the directions on an instrument are poorly understood, then scores may be affected. For example, questions in a self-report instrument may be interpreted differently by different respondents, leading to a distorted measure of the variable.
6. *Item sampling*. Errors can be introduced as a result of the sampling of items used in the measure. For example, a nursing student's score on a 100-item test of critical care nursing knowledge will be influenced by *which* 100 questions are included. A person might get 95 questions correct on one test but only 92 right on another similar test.
7. *Instrument format*. Technical characteristics of an instrument can influence measurements. For example, the ordering of questions in an instrument may influence responses.

TIP: The Toolkit section of Chapter 14 of the *Resource Manual* includes a list of suggestions for enhancing data quality and minimizing measurement error in quantitative studies.

RELIABILITY OF MEASURING INSTRUMENTS

The reliability of a quantitative instrument is a major criterion for assessing its quality. An instrument's **reliability** is the consistency with which it measures the target attribute. If a scale weighed a person at 120 pounds one minute and 150 pounds the next, it would be unreliable. The less variation an instrument produces in repeated measurements, the higher its reliability. Thus, reliability can be equated with a measure's stability, consistency, or dependability.

Reliability also concerns accuracy. An instrument is reliable to the extent that its measures reflect true scores—that is, to the extent that measurement errors are absent from obtained scores. Reliable measures maximize the true score component and minimize error.

These two ways of explaining reliability (consistency and accuracy) are not so different as they might appear. Errors of measurement that impinge on an instrument's accuracy also affect its consistency. The example of the scale with variable weight readings illustrates this point. Suppose that the true weight of a person is 125 pounds, but that two independent measurements yielded 120 and 150 pounds. In terms of the equation presented in the previous section, we could express the measurements as follows:

$$120 = 125 - 5$$

$$150 = 125 + 25$$

The errors of measurement for the two trials (–5 and +25, respectively) resulted in scores that are inconsistent *and* inaccurate.

The reliability of an instrument can be assessed in various ways, and the appropriate method depends on the nature of the instrument and on the aspect of reliability of greatest concern. Three key aspects are stability, internal consistency, and equivalence.

Stability

The **stability** of an instrument is the extent to which similar scores are obtained on separate occasions. The reliability estimate focuses on the instrument's susceptibility to extraneous influences over time, such as participant fatigue.

Assessments of stability involve procedures that evaluate **test–retest reliability**. Researchers administer the same measure to a sample twice and then compare the scores. The comparison is performed objectively by computing a **reliability coefficient**, which is an index of the magnitude of the test's reliability.

To explain reliability coefficients, we must discuss a statistic called a **correlation coefficient**. We have pointed out that researchers seek to detect and explain relationships among phenomena. For example, is there a relationship between patients' gastric acidity levels and degree of stress? The correlation

coefficient is a tool for quantitatively describing the magnitude and direction of a relationship between two variables. The computation of this index does not concern us here. It is more important to understand how to read a correlation coefficient.

Two variables that are obviously related are people's height and weight. Tall people tend to be heavier than short people. We would say that there was a **perfect relationship** if the tallest person in a population were the heaviest, the second tallest person were the second heaviest, and so forth. Correlation coefficients summarize how perfect a relationship is. The possible values for a correlation coefficient range from –1.00 through .00 to +1.00. If height and weight were perfectly correlated, the correlation coefficient expressing this relationship would be 1.00. Because the relationship exists but is not perfect, the correlation coefficient is in the vicinity of .50 or .60. The relationship between height and weight can be described as a **positive relationship** because *increases* in height tend to be associated with *increases* in weight.

When two variables are totally unrelated, the correlation coefficient equals zero. One might expect that women's dress sizes are unrelated to their intelligence. Large women are as likely to perform well on IQ tests as small women. The correlation coefficient summarizing such a relationship would presumably be in the vicinity of .00.

Correlation coefficients running from .00 to –1.00 express **inverse** or **negative relationships**. When two variables are inversely related, increases in one variable are associated with *decreases* in the second variable. Suppose that there is an inverse relationship between people's age and the amount of sleep they get. This means that, on average, the older the person, the fewer the hours of sleep. If the relationship were perfect (e.g., if the oldest person in a population got the least sleep, and so on), the correlation coefficient would be –1.00. In actuality, the relationship between age and sleep is probably modest—in the vicinity of –.15 or –.20. A correlation coefficient of this magnitude describes a weak relationship: older people *tend* to sleep fewer hours and younger people *tend* to sleep more, but nevertheless some younger people sleep few hours, and some older people sleep a lot.

Now, we can discuss the use of correlation coefficients to compute reliability estimates. With test–retest reliability, an instrument is administered twice to the same people. Suppose we wanted to assess the stability of a self-esteem scale. Self-esteem is a fairly stable attribute that does not fluctuate much from day to day, so we would expect a reliable measure of it to yield consistent scores on two occasions. To check the instrument's stability, we administer the scale 2 weeks apart to 10 people. Fictitious data for this example are presented in Table 14.1. It can be seen that, in general, differences in scores on the two testings are not large. The reliability coefficient for test–retest estimates is the correlation coefficient between the two sets of scores. In this example, the reliability coefficient is .95, which is high.

The value of the reliability coefficient theoretically can range between –1.00 and +1.00, like other correlation coefficients. A negative coefficient would have been obtained in our example if those with high self-esteem scores at time 1 had low scores at time 2, and vice versa. In practice, reliability coefficients usually range between .00 and 1.00. The higher the coefficient, the more stable the

TABLE 14.1 Fictitious Data for Test–Retest Reliability of Self-Esteem Scale

PARTICIPANT NUMBER	TIME 1	TIME 2	
1	55	57	
2	49	46	
3	78	74	
4	37	35	
5	44	46	
6	50	56	
7	58	55	
8	62	66	
9	48	50	
10	67	63	$r = .95$

measure. Reliability coefficients above .80 usually are considered good.

The test–retest method is easy, and can be used with self-report, observational, and physiologic measures. Yet, this approach has certain disadvantages. One issue is that many traits *do* change over time, independently of the measure's stability. Attitudes, knowledge, perceptions, and so on can be modified by experiences between testings. Test–retest procedures confound changes from measurement error with true changes in the attribute. Still, there are many relatively enduring attributes for which a test–retest approach is suitable.

Stability estimates suffer from other problems, however. One possibility is that people's responses (or observers' coding) on the second administration will be influenced by their memory of initial responses, regardless of the actual values the second day. Such memory interference results in spuriously high reliability coefficients. Another difficulty is that people may actually change *as a result of* the first administration. Finally, people may not be as careful using the same instrument a second time. If they find the process boring on the second occasion, then responses could be haphazard, resulting in a spuriously low estimate of stability.

On the whole, reliability coefficients tend to be higher for short-term retests than for long-term retests (those greater than 1 month) because of actual changes in the attribute being measured. Stability indexes are most appropriate for relatively stable characteristics such as personality, abilities, or certain physical attributes such as adult height.

It might be noted that while most test–retest efforts involve the calculation of a standard correlation coefficient, as just described, other methods are sometimes used. For example, Yen and Lo (2002) describe how an *intraclass correlation* (ICC) approach offers advantages because of the ability of this index to detect systematic error.

Example of test–retest reliability: Kao and Lynn (2009) developed the Family Caregiver Medication Administration Hassles Scale for use with Mexican American family caregivers of older relatives. The 3-week test–retest reliability for the scale was .64.

Internal Consistency

Scales and tests that involve summing item scores are typically evaluated for their internal consistency. Scales designed to measure an attribute ideally are composed of items that measure that attribute and nothing else. On a scale to measure nurses' empathy, it would be inappropriate to include an item that measures diagnostic competence. An instrument may be said to be **internally consistent** or *homogeneous* to the extent that its items measure the same trait.

Internal consistency reliability is the most widely used reliability approach. Its popularity reflects the fact that it is economical (it requires only one administration) and is the best means of assessing an especially important source of measurement error in psychosocial instruments, the sampling of items.

TIP: Many scales contain multiple **subscales**, each of which taps distinct but related concepts (e.g., a measure of fatigue might include subscales for mental and physical fatigue). The internal consistency of each subscale should be assessed. If subscale scores are summed for a total score, the scale's overall internal consistency is also computed.

The most widely used method for evaluating internal consistency is **coefficient alpha** (or **Cronbach's alpha**). Coefficient alpha can be interpreted like other reliability coefficients: the normal range of values is between .00 and +1.00, and higher values reflect higher internal consistency. It is beyond the scope of this text to explain this method in detail, but information is available in psychometric textbooks (e.g., Nunnally & Bernstein, 1994; Waltz, et al. 2010). Most statistical software can be used to calculate alpha. The research example at the end of Chapter 15 presents some computer output for a reliability analysis.

In summary, coefficient alpha is an index of internal consistency to estimate the extent to which different subparts of an instrument (i.e., items) are reliably measuring the critical attribute. Cronbach's alpha does not, however, evaluate fluctuations over time as a source of unreliability.

Example of internal consistency reliability: Villanueva and colleagues (2009) developed and evaluated a scale to measure nonpsychiatric healthcare providers' attitudes toward pediatric patients with mental illness. The 18-item scale had good internal consistency, alpha = .85.

Equivalence

Equivalence, in the context of reliability assessment, primarily concerns the degree to which two or more independent observers or coders agree about scoring. If there is a high level of agreement, then the assumption is that measurement errors have been minimized. Nurse researchers are especially likely to use this approach with observational measures, although it can be used in other applications—for example, for evaluating the consistency of coding open-ended questions or the accuracy of extracting data from records.

The reliability of ratings and classifications can be enhanced by careful training and the specification of clearly defined, nonoverlapping categories. Even when such care is taken, researchers should assess the reliability of observational instruments and coding systems. In this case, "instrument" includes both the category or rating system *and* the observers or coders making the measurements.

Interrater (or *interobserver*) **reliability** can be assessed using various approaches, which can be categorized as consensus, consistency, and measurement approaches (Stemler, 2004). Many interrater reliability indexes used by nurse researchers are of the consensus type, in which the goal is to have observers share a common interpretation of a construct, and to reach consensus (exact agreement). Consensus measures of interrater reliability for observational coding involve having two or more trained observers watching an event simultaneously, and independently recording data. The data are then used to compute an index of agreement between observers. (For coders, information would be independently coded into categories and then intercoder agreement would be assessed.) When ratings are dichotomous, one procedure is to calculate the proportion of agreements, using the following equation:

$$\frac{\text{Number of agreement}}{\text{Number of agreement} + \text{disagreements}}$$

This formula unfortunately tends to overestimate agreements because it fails to account for agreement by chance. If a behavior being observed were coded for absence versus presence, the observers would agree 50% of the time by chance alone. A widely used statistic in this situation is Cohen's **kappa**, which adjusts for chance agreements. Different standards have been proposed for acceptable levels of kappa, but there is some agreement that a value of .60 is minimally acceptable, and that values of .75 or higher are very good.

For certain types of data (e.g., ratings on a multipoint scale), correlation techniques are suitable, and these typically capture consistency rather than consensus. For example, a correlation coefficient can be computed to demonstrate the strength of the relationship between one rater's scores and another's. The **intraclass correlation coefficient** (ICC) can also be used to assess interrater reliability (Shrout & Fleiss, 1979).

Example of interrater reliability: Voepel-Lewis and colleagues (2010) assessed the FLACC Behavioral Scale, an observational tool to assess pain in critically ill patients. Exact agreement, kappa values, and intraclass correlation coefficients suggested strong interrater reliability of the measure.

Interpretation of Reliability Coefficients

Reliability coefficients are important indicators of an instrument's quality. Unreliable measures reduce statistical power and hence affect statistical conclusion validity. If data fail to support a hypothesis, one possibility is that the instruments were unreliable—not necessarily that the expected relationships do not exist. Knowing an instrument's reliability thus is critical in interpreting research results, especially if hypotheses are not supported.

For group-level comparisons, coefficients in the vicinity of .70 may be adequate (especially for

subscales), but coefficients of .80 or greater are highly desirable. By group-level comparisons, we mean that researchers compare scores of groups, such as male versus female or experimental versus control participants. The reliability coefficients for measures used for making decisions about individuals ideally should be .90 or better. For instance, if a test score was used as a criterion for admission to a nursing program, then the test's accuracy would be of critical importance to both the applicants and the school of nursing.

Reliability coefficients have a special interpretation that relates to our discussion of decomposing observed scores into error and true score components. Suppose we administered a scale that measures hopefulness to 50 patients with cancer. The scores would vary from one person to another—that is, some people would be more hopeful than others. Some variability in scores is true variability, reflecting real individual differences in hopefulness; some variability, however, is error. Thus,

$$V_O = V_T + V_E$$

where V_O = observed total variability in scores
V_T = true variability
V_E = variability owing to errors

A reliability coefficient is directly associated with this equation. *Reliability is the proportion of true variability to the total obtained variability,* or

$$r = \frac{V_T}{V_O}$$

If, for example, the reliability coefficient were .85, then 85% of the variability in obtained scores would represent true individual differences, and 15% of the variability would reflect extraneous fluctuations. Looked at in this way, it should be clear why instruments with reliability lower than .70 are risky to use.

Factors Affecting Reliability

Various things affect an instrument's reliability, and these factors are useful to keep in mind in selecting an instrument. First, the reliability of composite self-report and observational scales is partly a function of their length (i.e., number of items). To improve reliability, more items tapping the same concept should be added. Items that have no discriminating power (i.e., that elicit similar responses from everyone) should, however, be removed. Item analysis procedures for guiding decisions about item retention, modification, or deletion are outlined in Chapter 15.

With observational scales, reliability can be improved by greater precision in defining categories, or greater clarity in explaining the underlying construct for rating scales. The best means of enhancing reliability in observational studies, however, is thorough observer training.

An instrument's reliability is related in part to the heterogeneity of the sample with which it is used. The more homogeneous the sample (i.e., the more similar their scores), the lower the reliability coefficient will be. This is because instruments are designed to measure differences among those being measured. If the sample is homogeneous, then it is more difficult for the instrument to discriminate reliably among those who possess varying degrees of the attribute. For example, a depression scale will be less reliable when administered to a homeless sample than when it is used with a general population.

An instrument's reliability is not a fixed entity. *The reliability of an instrument is a property not of the instrument but rather of the instrument when administered to certain people under certain conditions*. A scale that reliably measures dependence in hospitalized adults may be unreliable with nursing homes residents. This means that in selecting an instrument, it is important to know the characteristics of the group with which it was developed. If the group is similar to the population for a new study, then the reliability estimate calculated by the scale developer is probably a reasonably good index of the instrument's accuracy in the new research.

TIP: You should not be satisfied with an instrument that will *probably* be reliable in your study. The recommended procedure is to compute new estimates of reliability whenever research data are collected.

Finally, reliability estimates vary according to the procedures used to obtain them. A scale's test–retest reliability is rarely the same value as its internal consistency reliability. In selecting an instrument, researchers need to determine which aspect of reliability (stability, internal consistency, or equivalence) is relevant.

Example of different reliability estimates: Schilling and colleagues (2009) developed a scale to measure self-management of type I diabetes among adolescents. They evaluated the scale's reliability using test–retest and internal consistency approaches. As an example of their findings, the coefficient alpha for the 7-item Goals subscale was .75. The subscale's test–retest reliability was .60 at 2 weeks and .59 at 3 months.

VALIDITY

A second key criterion for evaluating an instrument is its validity. **Validity** is the degree to which an instrument measures what it is supposed to measure. When researchers develop an instrument to measure hopelessness, they need to be sure that resulting scores validly reflect this construct and not something else, like depression.

Reliability and validity are not independent qualities of an instrument. *A measuring device that is unreliable cannot be valid*. An instrument cannot validly measure an attribute if it is inconsistent and inaccurate. An unreliable instrument contains too much error to be a valid indicator of the target variable. An instrument can, however, be reliable without being valid. Suppose we had the idea to assess patients' anxiety by measuring their height. We could obtain highly accurate, consistent measurements of their height, but such measures would not be valid indicators of anxiety. Thus, the high reliability of an instrument provides no evidence of its validity; low reliability *is* evidence of low validity.

Like reliability, validity has different aspects and assessment approaches, but unlike reliability, an instrument's validity is difficult to evaluate. There are no equations that can easily be applied to the scores of a hopelessness scale to estimate how good a job the scale is doing in measuring the critical variable. Validation is an evidence-building enterprise, in which the goal is to assemble sufficient evidence from which validity can be inferred. The greater the amount of evidence supporting validity, the more sound the inference.

TIP: Instrument developers usually gather evidence of the validity and reliability of their instrument in a **psychometric assessment** before making the instrument available for general use. If you use an existing instrument, choose one with demonstrated high reliability and validity.

Face Validity

Face validity refers to whether the instrument *looks* like it is measuring the target construct. Although face validity is not considered strong evidence of validity, it is helpful for a measure to have face validity if other types of validity have also been demonstrated. It might be easier to persuade people to participate in a study if the instruments have face validity, for example.

Example of face validity: Jones and colleagues (2008) developed the Stroke Self-Efficacy Questionnaire for use by practitioners working in stroke care. Face validity was addressed through consultation with experts in stroke rehabilitation and self-efficacy theory, as well as with stroke survivors.

Content Validity

Content validity concerns the degree to which an instrument has an appropriate sample of items for the construct being measured and adequately covers the construct domain. Content validity is relevant for both affective measures (i.e., measures of psychological traits) and cognitive measures.

For cognitive measures, the content validity question is, how representative are the test questions of the universe of questions on this topic? For example, suppose we were testing students' knowledge about major nursing theories. The test would

not be content valid if it omitted questions about, for example, Orem's Self-Care Theory.

Content validity is also relevant in developing affective measures. Researchers designing a new instrument should begin with a thorough conceptualization of the construct so the instrument can capture the full content domain. Such a conceptualization might come from a variety of sources, including rich first-hand knowledge, an exhaustive literature review, consultation with experts, or findings from a qualitative inquiry.

Example of using qualitative data to enhance content validity: Williams and Kristjanson (2009) developed a scale to measure hospitalized patients' perceptions of the emotional care they experienced. The items were based on the themes identified in a grounded theory study, which explored characteristics of interpersonal interactions patients perceived to be therapeutic.

An instrument's content validity is necessarily based on judgment. There are no completely objective methods of ensuring adequate content coverage on an instrument, but it is common to use a panel of experts to evaluate the content validity of new instruments.

There are various approaches to assessing content validity using an expert panel, but nurse researchers have been in the forefront in developing approaches that involve the calculation of a **content validity index (CVI)**. The experts are asked to evaluate individual items on the new measure as well as the overall instrument. Two key issues in such an evaluation are whether individual items are relevant and appropriate in terms of the construct, and whether the items taken together adequately measure all dimensions of the construct.

At the item level, a common procedure is to have experts rate items on a four-point scale of relevance. There are several variations of labeling the 4 points, but the scale used most often is as follows: 1 = *not relevant,* 2 = *somewhat relevant,* 3 = *quite relevant,* 4 = *highly relevant.* Then, for each item, the **item CVI (I-CVI)** is computed as the number of experts giving a rating of 3 or 4, divided by the number of experts—that is, the proportion in agreement about relevance. For example, an item rated as "quite" or "highly" relevant by 4 out of 5 judges would have an I-CVI of .80, which is considered an acceptable value.

There are two approaches to calculating **scale CVIs (S-CVIs)**, and unfortunately, instrument development papers seldom indicate which approach was used (Polit & Beck, 2006). One approach is to calculate the percentage of items on the scale for which *all* judges agreed on content validity. In other words, if a 10-item scale had 6 items for which the I-CVIs were 1.00, then the S-CVI would be .60. We call this the S-CVI/UA (universal agreement) approach. Because disagreements (as well as agreements) can occur by chance, and because disagreements could reflect bias or misunderstanding, we find this approach too stringent.

A second method is to compute the S-CVI by averaging I-CVIs. We recommend the averaging approach, which we refer to as S-CVI/Ave, and suggest a value of .90 as the standard for establishing excellent content validity (Polit & Beck, 2006). Content validation should be done with at least 3 experts, but a larger group is preferable. Further guidance is offered in Chapter 15.

Example of using a content validity index: Chien and Chan (2009) tested the Chinese version of the Level of Expressed Emotion Scale, a scale used with families of people with schizophrenia. The item-level CVIs ranged from .86 to 1.00 and the scale-level CVI, using the averaging approach, was .993.

Criterion-Related Validity

An instrument is said to have **criterion-related validity** if its scores correlate highly with scores on an external criterion. For example, if scores on a scale of attitudes toward premarital sex correlate highly with subsequent loss of virginity in a sample of teenagers, then the attitude scale would have good validity. For criterion-related validity, the key issue is whether the instrument is a useful predictor of other behaviors, experiences, or conditions.

A requirement of this approach is the availability of a reliable and valid criterion with which measures on the instrument can be compared. This is, unfortunately, seldom easy. If we were developing an instrument to measure nursing students' clinical skills, we might use supervisory ratings as our criterion—but can we be sure that these ratings are valid and reliable? The ratings might themselves need validation. Criterion-related validity is most appropriate when there is a concrete, reliable criterion. For example, a scale to measure smokers' motivation to quit smoking has a clear-cut, objective criterion: subsequent smoking.

Once a criterion is selected, a criterion-related **validity coefficient** can be computed by correlating scores on the instrument and the criterion. The magnitude of the coefficient is a direct estimate of how valid the instrument is, according to this validation method. To illustrate, suppose we developed a scale to measure nurses' professionalism. We administer the instrument to a sample of nurses and also ask the nurses to indicate how many professional conferences they have attended. The conference variable was chosen as one of many potential objective criteria of professionalism. Fictitious data are presented in Table 14.2. The correlation coefficient of .83 indicates that the professionalism scale correlates fairly well with the number of conferences attended. Whether the scale is really measuring professionalism is a different issue—an issue that is a construct validation concern discussed in the next section.

A distinction is sometimes made between two types of criterion-related validity. **Predictive validity** refers to the adequacy of an instrument in differentiating between people's performance on a future criterion. When a school of nursing correlates incoming students' high school grades with subsequent grade-point averages, the predictive validity of the high school grades for nursing school performance is being evaluated.

Example of predictive validity: Chang and colleagues (2009) developed and tested the Chinese version of the Positive and Negative Suicide Ideation Inventory. To assess predictive validity, a subsample of students used in the original instrument development study was recruited 1 year later to see if scores on the scale were predictive of recent suicide attempts.

Concurrent validity reflects an instrument's ability to distinguish individuals who differ on a present criterion. For example, a psychological test to differentiate between patients in a mental institution who can and cannot be released could be correlated with current behavioral ratings of healthcare

TABLE 14.2 Fictitious Data for Criterion-Related Validity Example

PARTICIPANT	SCORE ON PROFESSIONALISM SCALE	NUMBER OF NURSING CONFERENCES	
1	25	2	
2	30	4	
3	17	0	
4	20	1	
5	22	0	
6	27	2	
7	29	5	
8	19	1	
9	28	3	
10	15	1	$r = .83$

personnel. The difference between predictive and concurrent validity, then, is the difference in the timing of obtaining measurements on a criterion.

Example of concurrent validity: Cha and colleagues (2008) assessed the concurrent validity of a condom self-efficacy scale in Korean college students by correlating scores on the scale with actual condom use.

Criterion-related validation is most often used in practically oriented research. Criterion-related validity is helpful in assisting decision makers by giving them some assurance that their decisions will be effective, fair, and, in short, valid.

Construct Validity

Construct validity is a key criterion for assessing the quality of a study. As noted in Chapter 10, construct validity concerns inferences from study particulars (such as measures used to operationalize variables) to higher-order constructs. The key construct validity question in measurement is: What is this instrument *really* measuring? Unfortunately, the more abstract the concept, the more difficult it is to establish construct validity; at the same time, the more abstract the concept, the less suitable it is to rely on criterion-related validity. It is really not just a question of suitability, but feasibility. What objective criterion is there for such concepts as empathy or separation anxiety?

Construct validation of an instrument is a challenging but vital task. Construct validation is a hypothesis-testing endeavor, typically linked to a theoretical perspective about the construct. In validating a measure of death anxiety, its relationship to a criterion would be less informative than its correspondence to a cogent conceptualization of death anxiety. Construct validation can be approached in several ways, but it always involves logical analysis and hypothesis tests. Constructs are explicated in terms of other abstract concepts; researchers develop hypotheses about the manner in which the target construct functions in relation to other constructs.

There are a number of ways to gather evidence about construct validity, which we discuss in this section. It should also be noted, however, that if an instrument developer has taken strong steps to ensure the content validity of the instrument, construct validity will also be strengthened.

Known Groups

One construct validation approach is the **known-groups technique**, which yields evidence of **contrast validity**. In this procedure, the instrument is administered to groups hypothesized to differ on the critical attribute because of a known characteristic. For instance, in validating a measure of fear of childbirth, we could contrast the scores of primiparas and multiparas. We would expect that women who had never given birth would be more anxious than women who had done so, and so we might question the instrument's validity if such differences did not emerge. We would not necessarily expect large differences; some primiparas would feel little anxiety, and some multiparas would express fears. We would, however, hypothesize differences in *average* group scores.

Example of the known-groups technique: Gozum and Hacihasanoglu (2009) did a psychometric assessment of the Turkish version of the Medication Adherence Self-Efficacy Scale with a sample of hypertensive patients. Using the known-groups approach, they compared scale scores for those with controlled versus uncontrolled blood pressure.

Hypothesized Relationships

A similar method of construct validation involves testing hypothesized relationships, often on the basis of theory or prior research. This is really a variant of the known-groups approach, which involves hypotheses about the relationship between the measure of the construct and a variable representing group membership. A researcher might reason as follows:

- According to theory, construct X is positively related to construct Y.
- Instrument A is a measure of construct X; instrument B is a measure of construct Y.

- Scores on A and B are correlated positively, as predicted.
- Therefore, it is inferred that A and B are valid measures of X and Y.

This logical analysis does not constitute proof of construct validity, but yields important evidence. Construct validation is essentially an ongoing evidence-building enterprise.

Example of testing relationships: Simmons and colleagues (2009) developed and tested a scale to measure psychological adjustment in patients with an ostomy. In the construct validation efforts, they hypothesized that adjustment scores would be positively correlated with time elapsed since surgery and with scores on an acceptance of illness scale, and their hypotheses were supported.

Convergent and Discriminant Validity

The **multitrait–multimethod matrix method (MTMM)** is a significant construct validation tool (Campbell & Fiske, 1959). This procedure involves the concepts of convergence and discriminability. **Convergence** is evidence that different methods of measuring a construct yield similar results. Different measurement approaches should converge on the construct. **Discriminability** is the ability to differentiate the construct from other similar constructs. Campbell and Fiske argued that evidence of both convergence and discriminability should be brought to bear in construct validation.

To help explain the MTMM approach, fictitious data from a study to validate a "need for autonomy" measure are presented in Table 14.3. In using this approach, researchers must measure the critical concept by two or more methods. Suppose we measured need for autonomy in nursing home residents by (1) giving a sample of residents a self-report scale (the measure we are attempting to validate), (2) asking nurses to rate residents after observing them in a task designed to elicit autonomy or dependence, and (3) having residents react to a pictorial stimulus depicting an autonomy-relevant situation (a so-called *projective* measure).

A second requirement of the full MTMM is to measure a differentiating construct, using the same measuring methods. In the current example, suppose we wanted to differentiate "need for autonomy" from "need for affiliation." The discriminant concept must be similar to the focal concept, as in our example: We would expect that people with high need for autonomy would tend to be relatively low on need for affiliation. The point of including both concepts in a single validation study is to gather evidence

TABLE 14.3 Multitrait–Multimethod Matrix

		SELF-REPORT (1)		OBSERVATION (2)		PROJECTIVE (3)	
METHOD	**TRAITS**	**AUT_1**	**AFF_1**	**AUT_2**	**AFF_2**	**AUT_3**	**AFF_3**
Self-report (1)	AUT_1	(.88)					
	AFF_1	−.38	(.86)				
Observation (2)	AUT_2	.60	−.19	(.79)			
	AFF_2	−.21	.58	−.39	(.80)		
Projective (3)	AUT_3	.51	−.18	.55	−.12	(.74)	
	AFF_3	−.14	.49	−.17	.54	−.32	(.72)

AUT = need for autonomy trait; AFF = need for affiliation trait.

that the two concepts are distinct, rather than two different labels for the same underlying attribute.

The numbers in Table 14.3 represent correlation coefficients between scores on six measures (two traits × three methods). For instance, the coefficient of –.38 at the intersection of AUT_1–AFF_1 is the correlation between self-report scores on the need for autonomy and need for affiliation measures. Recall that a minus sign before the correlation coefficient signifies an inverse relationship. In this case, the –.38 tells us that there was a slight tendency for people scoring high on the need for autonomy scale to score low on the need for affiliation scale. (The numbers in parentheses along the diagonal of this matrix are the reliability coefficients.)

Various parts of the MTMM matrix have a bearing on construct validity. The most direct evidence (**convergent validity**) comes from the correlations between two different methods measuring the same trait. In the case of AUT_1–AUT_2, the coefficient is .60, which is reasonably high. Convergent validity should be large enough to encourage further scrutiny of the matrix. Second, the convergent validity entries should be higher, in absolute magnitude,* than correlations between measures that have neither method nor trait in common. That is, AUT_1–AUT_2 (.60) should be greater than AUT_2–AFF_1 (–.21) or AUT_1–AFF_2 (–.19), as it is here. This requirement is a minimum one that, if failed, should cause researchers to have serious doubts about the measures. Third, convergent validity coefficients should be greater than coefficients between measures of different traits by a single method. Once again, the matrix in Table 14.3 fulfills this criterion: AUT_1–AUT_2 (.60) and AUT_2–AUT_3 (.55) are higher in absolute value than AUT_1–AFF_1 (–.38), AUT_2–AFF_2 (–.39), and AUT_3–AFF_3 (–.32). The last two requirements provide evidence for **discriminant validity**.

The evidence is seldom as clear-cut as in this contrived example. Indeed, a common problem with MTMM is interpreting the pattern of coefficients. Another issue is that there are no clear-cut criteria for deciding whether MTMM requirements have been met—that is, there are no objective means of assessing the magnitude of similarities and differences within the matrix. The MTMM is nevertheless a valuable tool for exploring construct validity. Researchers sometimes decide to use MMTM concepts even when the full model is not feasible, as in focusing only on convergent validity.

Example of convergent and discriminant validity: Morea and colleagues (2008) developed and tested the Illness Self-Concept Scale, an instrument designed to predict adjustment in fibromyalgia. Their analyses provided some evidence that their construct, illness self-concept, is distinct from other similar constructs like depression (discriminant validity) and various analyses also supported evidence of convergent validity.

Factor Analysis

Another approach to construct validation uses a statistical procedure called factor analysis. Although factor analysis, which is discussed in Chapter 15, is computationally complex, it is conceptually rather simple. **Factor analysis** is a method for identifying clusters of related variables—that is, dimensions underlying a broad construct. Each dimension, or **factor**, represents a relatively unitary attribute. The procedure is used to identify and group together different items measuring an underlying attribute. In effect, factor analysis constitutes another means of testing hypotheses about the interrelationships among variables, and for looking at the convergent and discriminant validity of a large set of items. Indeed, a procedure known as **confirmatory factor analysis** (CFA) is sometimes used as a method for analyzing MTMM data (Ferketich, et al., 1991; Lowe & Ryan-Wenger, 1992).

Example of factor analysis in construct validation: Zheng and colleagues (2010) developed and tested the Dialysis Patient-Perceived Exercise Benefits and Barriers Scale. Responses to the scale's 24 items by a sample of 269 hemodialysis patients in China were factor analyzed to assess construct validity. Confirmatory factor analysis confirmed a 6-factor structure.

*__Absolute value__ refers to the value without a plus or minus sign. A value of −.80 is of a higher absolute magnitude than +.40.

Interpretation of Validity

Like reliability, validity is not an all-or-nothing characteristic of an instrument. An instrument does not possess or lack validity; it is a question of degree. An instrument's validity is not proved, established, or verified but rather is supported to a greater or lesser extent by evidence.

Strictly speaking, researchers do not validate an instrument but rather an application of it. A measure of anxiety may be valid for presurgical patients on the day of an operation but may not be valid for nursing students on the day of a test. Of course, some instruments may be valid for a wide range of uses with different types of samples, but each use requires new supporting evidence. The more evidence that can be gathered that an instrument is measuring what it is supposed to be measuring, the more confidence researchers will have in its validity.

TIP: When you select an instrument, you should seek evidence of the scale's psychometric soundness by examining the instrument developers' report. However, you also should consider evidence from others who have used the scale. Each time the scale "performs" as hypothesized, this constitutes supplementary evidence for its validity. Conversely, if hypotheses involving the use of the scale are not supported, this suggests potential validity problems (although, of course, other factors may account for nonsupported hypotheses, such as a small sample).

SENSITIVITY, SPECIFICITY, AND LIKELIHOOD RATIOS

Reliability and validity are the two most important criteria for evaluating quantitative instruments, but researchers sometimes need to consider other qualities of an instrument. In particular, sensitivity and specificity are criteria that are important in evaluating instruments used as screening or diagnostic tools (e.g., a scale to measure risk of osteoporosis). Screening/diagnostic instruments can be self-report, observational, or biophysiologic measures.

Sensitivity is the ability of a measure to identify a "case" correctly, that is, to screen in or diagnosis a condition correctly. A measure's sensitivity is its rate of yielding "true positives." **Specificity** is the measure's ability to identify noncases correctly, that is, to screen *out* those without the condition. Specificity is an instrument's rate of yielding "true negatives." To evaluate an instrument's sensitivity and specificity, researchers need a reliable and valid criterion of "caseness" against which scores on the instrument can be assessed.

Calculating Sensitivity, Specificity, and Related Indicators

Suppose we wanted to evaluate whether adolescents' self-reports about their smoking were accurate, and we asked 100 teenagers about whether they had smoked a cigarette in the previous 24 hours. The "gold standard" for nicotine consumption is cotinine levels in a body fluid, so assume that we did a urinary cotinine assay. Some fictitious data are shown in Table 14.4.

Sensitivity, in this example, is calculated as the proportion of teenagers who said they smoked *and* who had high concentrations of cotinine, divided by all real smokers as indicated by the urine test. Put another way, it is the true positives divided by all positives. In this case, there was considerable under-reporting of smoking and so the sensitivity of the self-report was only .50. Specificity is the proportion of teenagers who accurately reported they did not smoke, or the true negatives divided by all negatives. In our example, specificity is .83. There was considerably less over-reporting of smoking ("faking bad") than under-reporting ("faking good"). Sensitivity and specificity are often reported as percentages rather than proportions, by multiplying the proportions by 100.

Often, other related indicators are calculated with such data. **Predictive values** are posterior probabilities—the probability of an outcome after the results are known. A **positive predictive value** (or PPV) is the proportion of people with a positive result who have the target outcome or disease. In our example, the PPV is the proportion of teens who

TABLE 14.4 Example Illustrating Sensitivity, Specificity, and Likelihood Ratios

	URINARY COTININE LEVEL		
SELF-REPORTED SMOKING	**Positive (Cotinine > 200 ng/mL)**	**Negative (Cotinine ≤ 200 ng/mL)**	**Total**
Yes, smoked	A (true positive) 20	B (false positive) 10	A + B 30
No, did not smoke	C (false negative) 20	D (true negative) 50	C + D 70
Total	A + C 40	B + D 60	A + B + C + D 100

Sensitivity = A/(A + C)	= .50
Specificity = D/(B + D)	= .83
Positive predictive value (PPV) = A/(A + B)	= .67
Negative predictive value (NPV) = D/(C = D)	= .71
Likelihood ratio—positive (LR+) = sensitivity/(1 − specificity)	= 2.99
Likelihood ratio—negative (LR−) = (1 − sensitivity)/specificity	= .60

said they smoke who actually *do* smoke, according to the cotinine test results. Two out of three of those who reported smoking had high concentrations of cotinine, and so PPV = .67. A **negative predictive value** (NPV) is the proportion of people who have a negative test result who do not have the target outcome or disease. As shown in Table 14.4, 50 out of the 70 teenagers who reported not smoking actually were nonsmokers, and so NPV in our example is .71.

Example of sensitivity, specificity, and predictive values: Chichero and colleagues (2009) developed a dysphagia screening tool to triage patients at risk of dysphagia on admission to acute hospital wards. Sensitivity was 95% and specificity was 97%. Positive predictive value was 92% and negative predictive value was 98%.

In the medical community, reporting **likelihood ratios** has come into favor because it summarizes the relationship between specificity and sensitivity in a single number. The likelihood ratio addresses the question, "How much more likely are we to find that an indicator is positive among those *with* the outcome of concern compared to those for whom the indicator is negative?" For a positive test result, then, the likelihood ratio (LR+) is the ratio of true-positive results to false-positive results. The formula for LR+ is sensitivity divided by 1 minus specificity. For the data in Table 14.4, LR+ is 2.99: We are about three times as likely to find that a self-report of smoking really *is* for a true smoker than it is for a nonsmoker. For a negative test result, the likelihood ratio (LR−) is the ratio of false-negative results to true-negative results. For the data in Table 14.4, the LR− is .60. In our example, we are about half as likely to find that a self-report of nonsmoking is false than we are to find that it reflects a true nonsmoker. When a test is high on both sensitivity and specificity (which is not especially true in our example), the likelihood ratio is high and discrimination is good.

Example of likelihood ratios: Novotny and Anderson (2008) tested an algorithm for predicting the probability of readmission (Pra) of medical inpatients within 41 days of discharge from the hospital, using hospital records data. Pra score values ranged from .16 to .75. With a Pra value of .45, the likelihood ratio was 1.6.

Receiver Operating Characteristic (ROC) Curves

All of the indicators that we calculated for the data in Table 14.4 are contingent upon the critical value that we established for cotinine concentration. Sensitivity and specificity would be quite different if we had used 100 ng/mL as indicative of smoking status, rather than 200 ng/mL. There is almost invariably a trade-off between the sensitivity and specificity of a measure. When sensitivity is increased to include more true positives, the proportion of true negatives declines. Therefore, a critical task in developing new diagnostic or screening measures is to develop the appropriate **cutoff point** (or *cutpoint*), that is, a score to distinguish cases and noncases.

To identify the best cutoff point, researchers often are guided by a **receiver operating characteristic curve** (**ROC curve**) (Fletcher, et al., 2005). To construct an ROC curve, the sensitivity of an instrument (i.e., the rate of correctly identifying a case vis-à-vis a well-established criterion) is plotted against the false-positive rate (i.e., the rate of incorrectly diagnosing someone as a case, which is the inverse of its specificity) over a range of different scores. The score (cutoff point) that yields the best balance between sensitivity and specificity can then be determined. The optimum cutoff is at or near the shoulder of the ROC curve.

ROC curves can best be explained with an illustration. Figure 14.2 presents an ROC curve from a study in which a goal was to establish cutoff points for scores on the Braden Q scale for predicting pressure ulcer risk in children (Curley et al., 2003). In this figure, sensitivity and one minus specificity are plotted for each possible score of the Braden Q scale. The upper left corner represents sensitivity at its highest possible value (1.0) and false positives at its lowest possible value (.00). Screening instruments that do an excellent job of discriminating

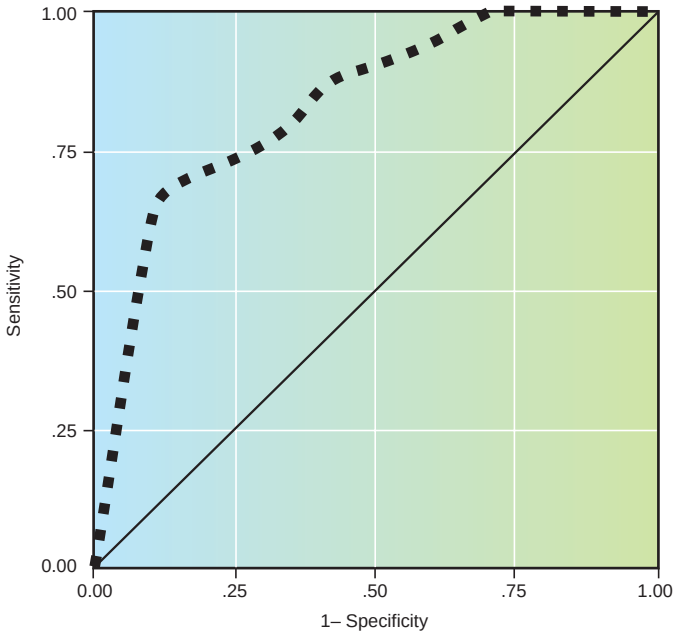


FIGURE 14.2 Receiver operating characteristic (ROC) curve for Braden Q Scale. From Curley, M. A. Q., Razmus, I. S., Roberts, K. E., & Wypij, D. (2003). Predicting pressure ulcer risk in pediatric patients: The Braden Q Scale. *Nursing Research, 52*, p. 27.

have points that crowd close to the upper left corner, which indicates that as sensitivity increases there is relatively little loss in specificity. ROC curves that are closer to a diagonal, from lower left to upper right, are indicative of an instrument with poor discriminatory power.

The overall accuracy of an instrument can be calculated as the proportion of the area under the ROC curve, an index referred to as **area under the curve**, or **AUC**. The larger the area, the more accurate the instrument. The AUC for the data portrayed in Figure 14.2 is .83. The cutoff score in this example was established at 16. At this cutoff value, the sensitivity was .88 and the specificity was .58. The researchers used these preliminary analyses to improve on the Braden Q scale and achieved even better results.

In selecting an appropriate cutoff point, the final decision is likely to be driven by clinical or economic factors and not just statistical ones. The financial and emotional costs of misclassifying people may be greater for false positives than false negatives, or vice versa.

OTHER CRITERIA FOR ASSESSING QUANTITATIVE MEASURES

Although we have already discussed the major criteria that are used to evaluate the quality of measuring instruments, we briefly mention a few others.

Efficiency

Instruments of comparable reliability and validity may differ in their efficiency. A depression scale that requires 5 minutes of people's time is efficient compared with a depression scale that requires 20 minutes to complete. In most studies, efficient instruments are desirable because they reduce participant burden.

One aspect of efficiency is the number of items on the instrument. Long instruments tend to be more reliable than shorter ones, but there is a point of diminishing returns. As an example, consider a 40-item scale to measure social support that has an internal consistency reliability of .94. We can use a formula, known as the **Spearman-Brown formula**, to estimate how reliable the scale would be with fewer items. As an example, if we wanted to shorten the scale to 30 items, the formula would result in an estimated reliability of .92.** Thus, a 25% reduction in the instrument's length resulted in a negligible decrease in reliability, from .94 to .92. Most researchers likely would sacrifice a modest amount of reliability in exchange for reducing response burden and data collection costs. Other things being equal, it is desirable to select as efficient an instrument as possible.

Other Criteria

A few remaining qualities that sometimes are considered in assessing a quantitative instrument can be noted. Most of the following criteria are actually aspects of the reliability and validity:

1. *Comprehensibility*. Participants and researchers should be able to comprehend the behaviors required to secure accurate and valid measures.
2. *Precision*. An instrument should discriminate between people with different amounts of an attribute as precisely as possible.
3. *Range*. The instrument should be capable of achieving a meaningful measure from the smallest expected value of the variable to the largest.
4. *Linearity*. A researcher normally strives to construct measures that are equally accurate and sensitive over the entire range of values.
5. *Reactivity*. The instrument should, insofar as possible, avoid affecting the attribute being measured.

**The equation (and the worked-out example) for this situation is as follows:

$$r^1 = \frac{kr}{1 + [(k - 1)r]} = \frac{.75(.94)}{1 + [(-.25)(.94)]} = .92$$

where k = the factor by which the instrument is being decreased, in this case, $k = 30 \div 40 = .75$; r = reliability for the full scale, here, .94; and r^1 = reliability estimate for the shorter scale.

DATA QUALITY WITH SINGLE INDICATORS

The discussion in this chapter has primarily focused on methods of evaluating data quality for multi-item scales, which are widely used by nurse researchers. Textbooks on research methods or measurement rarely say much about reliability or validity for single questions (e.g., "What is your date of birth?") or single-item scales, such as visual analog scales.

The truth of the matter is that it is not easy to evaluate data quality in such situations. This is of great concern in large national surveys, such as the National Longitudinal Study of Adolescent Health. Population estimates of, say, average number of times adolescents have been hospitalized, or the percentage who have ever used marijuana, are based on reports in response to individual (nonscaled) questions, so the accuracy of the responses is vital. We touch briefly here on data quality assessment for single indicators.

The two basic strategies for estimating measurement error in such situations are a test–retest approach and external verification. In the former, the questions that are of interest are asked on two separate occasions. When this happens for the express purpose of assessing consistency (in what is called a *response variance reinterview*), the second administration typically involves a subsample of respondents and an abbreviated instrument with key questions. Survey researchers compute various statistical indexes (e.g., an *index of inconsistency*) to help them understand and interpret response differences—that is, measurement error—in the two administrations (Subcommittee on Measuring and Reporting the Quality of Survey Data, 2001). Although few nurse researchers would have the resources to undertake such an enterprise, there may be opportunities to use the underlying principle for critical pieces of information. For example, in a self-report instrument, it might be possible to ask the same question twice, early and later, for example, or to ask the question in slightly different ways in the same questionnaire or interview. Also, if a study is longitudinal, factual information (e.g., date of birth) could be gathered twice to assess any discrepancies.

The second approach is to verify information provided in the primary data gathering method against an external source—a form of criterion-related validation. For example, information from a question about birth date could be checked against birth records. Responses to questions about health status, diagnosis, or healthcare could be checked against medical records. Measurement errors are then estimated based on a comparison of the two types of information. It should not necessarily be assumed that records are free of error, but they may be less prone to certain types of bias. Other forms of external verification may be available. In particular, **proxy reports** (obtaining data from another person, such as a family member) might be an option. Patrician (2004) has offered additional guidance regarding single-item scales.

Researchers using biophysiologic measures should also give data quality some thought rather than assuming they will be error free. Instruments may not be properly calibrated, the person doing the tests may not follow laboratory protocols, and laboratory procedures can vary from one lab to the next. Measurement errors can also occur because of patient circumstances, such as insufficient sleep. Moreover, if physiologic measures are taken from charts, the possibility of error should be considered.

CRITIQUING DATA QUALITY IN QUANTITATIVE STUDIES

If data are seriously flawed, the study cannot contribute useful evidence. Therefore, in drawing conclusions about a study's evidence, it is important to consider whether researchers have taken appropriate steps to collect data that accurately reflect reality. Research consumers have the right—indeed, the obligation—to ask: Can I trust the data? Do the data accurately and validly reflect key constructs?

Information about data quality should be provided in every quantitative research report because it is not possible to come to conclusions about the quality of study evidence without such information. Reliability estimates are usually reported because they are

BOX 14.1 Guidelines for Critiquing Data Quality in Quantitative Studies

1. Is there congruence between the research variables as conceptualized (i.e., as discussed in the introduction of the report) and as operationalized (i.e., as described in the method section)?
2. If operational definitions (or scoring procedures) are specified, do they clearly indicate the rules of measurement? Do the rules seem sensible? Were data collected in such a way that measurement errors were minimized?
3. Does the report offer evidence of the reliability of measures? Does the evidence come from the research sample itself, or is it based on other studies? If the latter, is it reasonable to conclude that data quality would be similar for the research sample as for the reliability sample (e.g., are sample characteristics similar)?
4. If reliability is reported, which estimation method was used? Was this method appropriate? Should an alternative or additional method of reliability appraisal have been used? Is the reliability sufficiently high?
5. Does the report offer evidence of the validity of the measures? Does the evidence come from the research sample itself, or is it based on other studies? If the latter, is it reasonable to believe that data quality would be similar for the research sample as for the validity sample (e.g., are the sample characteristics similar)?
6. If validity information is reported, which validity approach was used? Was this method appropriate? Does the validity of the instrument appear to be adequate?
7. If there is no reliability or validity information, what conclusion can you reach about the quality of the data in the study?
8. If a diagnostic or screening tool was used, is information provided about its sensitivity and specificity, and were these qualities adequate?
9. Were the research hypotheses supported? If not, might data quality play a role in the failure to confirm the hypotheses?

easy to communicate. Ideally—especially for composite scales—the report should provide reliability coefficients based on data from the study itself, not just from previous research. Interrater or interobserver reliability is especially crucial for coming to conclusions about data quality in observational studies. The values of the reliability coefficients should be sufficiently high to support confidence in the findings. It is especially important to scrutinize reliability information in studies with nonsignificant findings because the unreliability of measures can undermine statistical conclusion validity.

Validity is more difficult to document in a report than reliability. At a minimum, researchers should defend their choice of existing measures based on validity information from the developers, and they should cite the relevant publication. If a study used a screening or diagnostic measure, information should also be provided about its sensitivity and specificity.

Box 14.1 provides some guidelines for critiquing aspects of data quality of quantitative measures. The guidelines are available in the Toolkit of the accompanying *Resource Manual* for your use and adaptation.

RESEARCH EXAMPLE

In this section, we describe a study that used both self-report and observational measures. We focus on the researchers' excellent documentation of data quality in their study.

Study: Communication and outcomes of visits between older patients and nurse practitioners (Gilbert and Hayes, 2009)

Statement of Purpose: The purpose of this study was to examine relationships among patient–clinician communication, background characteristics of the patients and the clinicians (nurse practitioners or NPs), and both proximal outcomes (e.g., patient satisfaction) and longer-term outcomes (e.g., changes in patients' physical and mental health).

Design: Visits between 31 NPs and 155 patients were video recorded and various aspects of patient and NP behaviors were coded. Proximal outcomes were measured by self-report after the visits. Four weeks later, changes in patients' health outcomes were assessed using self-report measures.

Instruments and Data Quality: Communications during the visits were measured using the Roter Interaction Analysis System (RIAS) for verbal interaction and a checklist for nonverbal behaviors. The Roter system involves coding for both the content of the communication and relationship aspects, using a system of 69 categories for all utterances (only 43 were used in this study). The researchers noted that the predictive validity of the RIAS had considerable support. The average interrater reliability in the present study for the 43 coded behavior categories was .95. For the nonverbal behavior checklist, various actions (e.g., gazes, nods, smiles) were coded in 1-second segments over a 30-second sample. Two coders independently coded all segments and any discrepancies in coding were resolved by a third party. Several variables were measured by patients' self-report, including both 1-item measures (e.g., satisfaction with the visit) and multi-item scales (e.g., physical and mental health). For example, patient satisfaction with the NP visit was measured using one item, previously used in a large national survey, which asked for ratings of perceived quality of care on a 10-point scale from 1 (*worst care possible*) to 10 (*best care possible*). The authors noted that a correlation of .72 between the ratings and the average of several other satisfaction items provided some evidence for the reliability of the single item. Physical and mental health were measured with a 12-item scale called the SF-12 Health Survey, a widely used and well-validated instrument. The test developer had reported results indicating Cronbach alpha values of .89 for physical health and .82 for mental health among people 65 years and older. In the present study, the researchers computed the internal consistency reliability to be .87 and .72 for physical and mental health, respectively.

Key Findings: Among the many findings reported in this study, the researchers found that better patient outcomes were associated with a higher amount of communication content involving seeking and giving biomedical and psychosocial information, and with a relationships component of more positive talk and greater trust and receptivity.

SUMMARY POINTS

- **Measurement** involves assigning numbers to objects to represent the amount of an attribute, using a specified set of rules. Researchers strive to develop or use measurements whose rules are *isomorphic* with reality.
- Few quantitative measuring instruments are infallible. Sources of measurement error include situational contaminants, response-set biases, and transitory personal factors, such as fatigue.
- **Obtained scores** from an instrument consist of a **true score** component (the value that would be obtained for a hypothetical perfect measure of the attribute) and an error component, or **error of measurement**, that represents measurement inaccuracies.
- **Reliability**, one of two primary criteria for assessing an instrument, is the degree of consistency or accuracy with which an instrument measures an attribute. The higher an instrument's reliability, the lower the amount of error in obtained scores.
- There are different methods for assessing an instrument's reliability and for computing a **reliability coefficient**. A reliability coefficient typically is based on the computation of a **correlation coefficient** that indicates the magnitude and direction of a relationship between two variables.
- Correlation coefficients can range from –1.00 (a **perfect negative relationship**) through zero to +1.00 (a **perfect positive relationship**). Reliability coefficients usually range from .00 to 1.00, with higher values reflecting greater reliability.
- The **stability** aspect of reliability, which concerns the extent to which an instrument yields the same results on repeated administrations, is evaluated as **test–retest reliability**.
- The **internal consistency** aspect of reliability—the extent to which all the instrument's items are measuring the same attribute—is usually assessed by **Cronbach's alpha**.
- When the reliability assessment focuses on **equivalence** between observers in rating or coding behaviors, estimates of **interrater** (or

interobserver) **reliability** are obtained. When a consensus measure capturing interrater agreement within a small number of categories is desired, the **kappa** statistic is often used.

- **Reliability coefficients** reflect the proportion of true variability in a set of scores to the total obtained variability.
- **Validity** is the degree to which an instrument measures what it is supposed to measure.
- **Face validity** refers to whether the instrument appears, on the face of it, to be measuring the appropriate construct.
- **Content validity** concerns the sampling adequacy of the content being measured. Expert ratings on the relevance of items can be used to compute **content validity index** (**CVI**) information. **Item CVIs** (**I-CVIs**) represent the proportion of experts rating each item as relevant. A **scale CVI** using the averaging calculation method (S-CVI/Ave**)** is the average of all I-CVI values.
- **Criterion-related validity** (which includes both **predictive validity** and **concurrent validity**) focuses on the correlation between the instrument and an outside criterion.
- **Construct validity**, an instrument's adequacy in measuring the focal construct, is a hypothesis-testing endeavor. One approach assesses **contrast validity**, using the **known-groups technique** to contrast scores of groups hypothesized to differ on the attribute; another approach is **factor analysis**, a statistical procedure for identifying unitary clusters of items or measures.
- Another construct validity approach is the **multitrait–multimethod (MTMM) matrix technique**, which is based on the concepts of convergence and discriminability. **Convergence** refers to evidence that different methods of measuring the same attribute yield similar results. **Discriminability** refers to the ability to differentiate the construct being measured from other, similar concepts.
- A **psychometric assessment** of a new instrument is usually undertaken to gather evidence about validity, reliability, and other assessment criteria.
- Sensitivity and specificity are important criteria for screening and diagnostic instruments. **Sensitivity** is the instrument's ability to identify a case correctly (i.e., its rate of yielding true positives). **Specificity** is the instrument's ability to identify noncases correctly (i.e., its rate of yielding true negatives). Other related indexes include the measure's **positive predictive value (PPV), negative predictive value (NPV)**, and **likelihood ratios**.
- Sensitivity is sometimes plotted against specificity in a **receiver operating characteristic curve** (**ROC curve**) to determine the optimum **cutoff point** for caseness.

STUDY ACTIVITIES

Chapter 14 of the *Resource Manual for Nursing Research*: *Generating and Assessing Evidence for Nursing Practice, 9th edition*, offers exercises and study suggestions for reinforcing concepts presented in this chapter. In addition, the following study questions can be addressed:

1. Explain in your own words the meaning of the following correlation coefficients:
   a. The relationship between intelligence and grade-point average was found to be .72.
   b. The correlation coefficient between age and gregariousness was –.20.
   c. It was revealed that patients' compliance with nursing instructions was related to their length of stay in the hospital ($r = -.50$).
2. Use the critiquing guidelines in Box 14.1 to evaluate data quality in the study by Gilbert and Hayes (2009), referring to the original study if possible.

STUDIES CITED IN CHAPTER 14

Cha, E., Kim, K., & Burke, L. (2008). Psychometric validation of a condom self-efficacy scale in Korean. *Nursing Research, 57*, 245–251.

Chang, H., Lin, C., Chou, K., Ma, W., & Yang, C. (2009). Chinese version of the positive and negative suicide ideation: Instrument development. *Journal of Advanced Nursing, 65*, 1485–1496.

Chichero, J., Heaton, S., & Bassett, L. (2009). Triaging dysphagia: Nurses screening for dysphagia in an acute hospital. *Journal of Clinical Nursing, 18*, 1649–1659.

Chien, W. T., & Chan, S. (2009). Testing the psychometric properties of a Chinese version of the Level of Expressed Emotion Scale. *Research in Nursing & Health, 32*, 59–70.

Curley, M. A. Q., Razmus, I. S., Roberts, K. E., & Wypij, D. (2003). Predicting pressure ulcer risk in pediatric patients. *Nursing Research, 52*, 22–33.

Gilbert, D., & Hayes, E. (2009). Communication and outcomes of visits between older patients and nurse practitioners. *Nursing Research, 58*, 283–293.

Gozum, S., & Hacihasanoglu, R. (2009). Reliability and validity of the Turkish adaptation of Medication Adherence Self-Efficacy Scale in hypertensive patients. *European Journal of Cardiovascular Nursing, 8*, 129–136.

Jones, F., Partridge, C., & Reid, F. (2008). The Stroke Self-Efficacy Questionnaire: Measuring individual confidence in functional performance after stroke. *Journal of Clinical Nursing, 17*, 244–252.

Kao, H., & Lynn, M. (2009). Use of the measurement of medication administration hassles with Mexican American family caregivers. *Journal of Clinical Nursing, 18*, 2596–2603.

Morea, J., Friend, R., & Bennett, R. (2008). Conceptualizing and measuring illness self-concept. A comparison with self-esteem and optimism in predicting fibromyalgia adjustment. *Research in Nursing & Health, 31*, 563–575.

Novotny, N., & Anderson, M. A. (2008). Prediction of early readmission in medical inpatients using the probability of repeated admission instrument. *Nursing Research, 57*, 406–415.

Schilling, L., Dixon, J., Knafl, K., Lynn, M., Murphy, K., Dumser, S., & Grey, M. (2009). A new self-report measure of self-management of type I diabetes for adolescents. *Nursing Research, 58*, 228–236.

Simmons, K., Smith, J., & Maekawa, A. (2009). Development and psychometric evaluation of the Ostomy Adjustment Inventory-23. *Journal of Wound, Ostomy & Continence Nursing, 36*, 69–75.

Williams, A., & Kristjanson, L. (2009). Emotional care experienced by hospitalized patients: Development and testing of a measurement instrument. *Journal of Clinical Nursing, 18*, 1069–1077.

Villanueva, C., Scott, S., Guzzetta, C., & Foster, B. (2009). Development and psychometric testing of the Attitudes toward Mental Illness in Pediatric Patients Scale. *Journal of Child & Adolescent Psychiatric Nursing, 22*, 220–227.

Voepel-Lewis, T., Zanotti, J., Dammeyer, J., & Merkel, S. (2010). Reliability and validity of the Face, Legs, Activity, Cry, Consolability Behavioral Tool in assessing acute pain in critically ill patients. *American Journal of Critical Care, 19*, 55–61.

Zheng, J., You, L., Lou, T., Chen, N., Lai, D., Liang, Y., Li, Y. N., Gu, Y. M., Lv, S. F., & Zhai, C. Q. (2010). Development and psychometric evaluation of the Dialysis Patient-Perceived Exercise Benefits and Barriers Scale. *International Journal of Nursing Studies, 47*, 166–180.

Methodologic and nonresearch references cited in this chapter can be found in a separate section at the end of the book.

7 Ethics in Nursing Research

In studies involving human beings or animals, researchers must deal with ethical issues. Ethics can be challenging because ethical requirements sometimes conflict with the desire to produce rigorous evidence. This chapter discusses major ethical principles that must be considered in designing research.

ETHICS AND RESEARCH

When humans are used as study participants, care must be exercised to ensure that their rights are protected. Ethical research conduct may strike you as self-evident, but ethical considerations have not always been given adequate attention.

Historical Background

The Nazi medical experiments of the 1940s are a famous example of disregard for ethical conduct. Nazi research involved the use of prisoners of war and racial "enemies" in experiments testing human endurance and reaction to untested drugs. The studies were unethical not only because they exposed people to harm and even death, but also because people could not refuse participation. Similar wartime experiments that raised ethical concerns were conducted in Japan and Australia (McNeill, 1993).

More recently, researchers investigated the effects of syphilis among poor African American men between 1932 and 1972 in the Tuskegee Syphilis Study, sponsored by the U.S. Public Health Service. Medical treatment was deliberately withheld to study the course of the untreated disease. A public health nurse recruited many participants (Vessey and Gennarao, 1994). Similarly, Dr. Herbert Green studied women with cervical cancer in Auckland, New Zealand in the 1980s; patients with carcinoma were not given treatment so that the natural progression of the disease could be studied.

In the Willowbrook Study, Dr. Saul Krugman conducted research on hepatitis during the 1960s. At Willowbrook, an institution for the mentally retarded on Staten Island, children were deliberately infected with the hepatitis virus. Even more recently, it was revealed in 1993 that U.S. federal agencies had sponsored radiation experiments since the 1940s on hundreds of people, many of them prisoners or elderly hospital patients. And in 2010, it was revealed that a U.S. doctor who worked on the Tuskegee Study inoculated prisoners in Guatemala with syphilis in the 1940s (Reverby, in press). Many other examples of studies with ethical transgressions—often more subtle than these examples—have emerged to give ethical concerns the high visibility they have today.

Codes of Ethics

In response to human rights violations, various **codes of ethics** have been developed. The *Nuremberg Code*, developed after Nazi atrocities were made public in the Nuremberg trials, was an international effort to establish ethical standards. The *Declaration of Helsinki*, another international set of standards, was adopted in 1964 by the World Medical Association and was most recently revised in 2008.

Most disciplines (e.g., psychology, sociology, medicine) have established their own ethical codes. In nursing, the American Nurses Association (ANA) issued *Ethical Guidelines in the Conduct, Dissemination, and Implementation of Nursing Research* (Silva, 1995). ANA also published in 2001 a revised *Code of Ethics for Nurses with Interpretive Statements*, a document that covers primarily ethical issues for practicing nurses but that also includes principles that apply to nurse researchers. In Canada, the Canadian Nurses Association published a document entitled *Ethical Research Guidelines for Registered Nurses* in 2002. In Australia, three nursing organizations collaborated to develop the *Code of Ethics for Nurses in Australia* (2008).

Some nurse ethicists have called for an international ethics code for nursing, but nurses in most countries have developed their own professional codes or follow the codes established by their governments. The International Council of Nurses (ICN), however, has developed the *ICN Code of Ethics for Nurses*, updated in 2006.

TIP: In their study of 27 ethical review boards in the United States, Rothstein & Phuong (2007) found nurses to be more sensitive to ethical issues than members from other disciplines.

Government Regulations for Protecting Study Participants

Governments throughout the world fund research and establish rules for adhering to ethical principles. For example, Health Canada specified the *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans* as the guidelines to protect study participants in all types of research. In Australia, the National Health and Medical Research Council issued the *National Statement on Ethical Conduct in Research Involving Humans* in 2007 and also issued a special statement about incentive payments to study participants in 2009.

In the United States, the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research adopted a code of ethics in 1978. The commission, established by the National Research Act, issued the ***Belmont Report***, which provided a model for many disciplinary guidelines. The Belmont Report also served as the basis for regulations affecting research sponsored by the U.S. government, including studies supported by NINR. The U.S. Department of Health and Human Services (DHHS) has issued ethical regulations that have been codified as Title 45 Part 46 of the Code of Federal Regulations (45 CFR 46). These regulations, revised most recently in 2005, are among the most widely used guidelines in the United States for evaluating the ethical aspects of studies.

TIP: There are many useful websites devoted to ethical principles, only some of which are mentioned in this chapter. Several websites are listed in the "Useful Websites for Chapter 7" file in the Toolkit of the accompanying *Resource Manual*, for you to click on directly.

Ethical Dilemmas in Conducting Research

Research that violates ethical principles is rarely done specifically to be cruel, but usually occurs out of a conviction that knowledge is important and potentially beneficial in the long run. There are situations in which participants' rights and study demands are in direct conflict, posing **ethical dilemmas** for researchers. Here are examples of research problems in which the desire for rigor conflicts with ethical considerations:

1. *Research question:* Are nurses equally empathic in their treatment of male and female patients in the ICU?
   Ethical dilemma: Ethics require that participants be aware of their role in a study. Yet if the researcher informs nurse participants that their empathy in treating male and female ICU patients will be scrutinized, will their behavior be "normal?" If the nurses' usual behavior is altered because of the known presence of research observers, then the findings will be inaccurate.
2. *Research question:* What are the coping mechanisms of parents whose children have a terminal illness?
   Ethical dilemma: To answer this question, the researcher may need to probe into the psychological state of parents at a vulnerable time; such probing could be painful or traumatic. Yet knowledge of the parents' coping mechanisms might help to design effective interventions for dealing with parents' grief and stress.
3. *Research question:* Does a new medication prolong life in patients with cancer?
   Ethical dilemma: The best way to test the effectiveness of an intervention is to administer the intervention to some participants but withhold it from others to see if differences between the groups emerge. However, if the intervention is untested (e.g., a new drug), the group receiving the intervention may be exposed to potentially hazardous side effects. On the other hand, the group *not* receiving the drug may be denied a beneficial treatment.
4. *Research question:* What is the process by which adult children adapt to the day-to-day stresses of caring for a parent with Alzheimer's disease?
   Ethical dilemma: Sometimes, especially in qualitative studies, a researcher may get so close to participants that they become willing to share "secrets" and privileged information. Interviews can become confessions—sometimes of unseemly or even illegal behavior. In this example, suppose a woman admitted to physically abusing her mother—how does the researcher respond to that information without undermining a pledge of confidentiality? And, if the researcher divulges the information to authorities, how can a pledge of confidentiality be given in good faith to other participants?

As these examples suggest, researchers are sometimes in a bind. Their goal is to develop high-quality evidence for practice, using the best methods available, but they must also adhere to rules for protecting human rights. Another dilemma can arise if nurse researchers are confronted with conflict-of-interest situations, in which their expected behavior as researchers conflicts with their expected behavior as nurses (e.g., deviating from a research protocol to give assistance to a patient). It is precisely because of such conflicts and dilemmas that codes of ethics have been developed to guide researchers' efforts.

ETHICAL PRINCIPLES FOR PROTECTING STUDY PARTICIPANTS

The *Belmont Report* articulated three broad principles on which standards of ethical conduct in research are based: beneficence, respect for human dignity, and justice. We briefly discuss these principles and then describe procedures researchers adopt to comply with them.

Beneficence

Beneficence imposes a duty on researchers to minimize harm and maximize benefits. Human research should be intended to produce benefits for participants or—a situation that is more common—for others. This principle covers multiple dimensions.

The Right to Freedom from Harm and Discomfort

Researchers have an obligation to avoid, prevent, or minimize harm (*nonmaleficence*) in studies with humans. Participants must not be subjected to unnecessary risks of harm or discomfort, and their participation must be essential to achieving scientifically and societally important aims that could

not otherwise be realized. In research with humans, *harm* and *discomfort* can be physical (e.g., injury, fatigue), emotional (e.g., stress, fear), social (e.g., loss of social support), or financial (e.g., loss of wages). Ethical researchers must use strategies to minimize all types of harms and discomforts, even ones that are temporary.

Research should be conducted only by qualified people, especially if potentially dangerous equipment or specialized procedures are used. Ethical researchers must be prepared to terminate a study if they suspect that continuation would result in injury, death, or undue distress to participants. When a new medical procedure or drug is being tested, it is usually advisable to experiment with animals or tissue cultures before proceeding to tests with humans. (Guidelines for the ethical treatment of animals are discussed later in this chapter.)

Protecting human beings from physical harm may be straightforward, but the psychological consequences of study participation are usually subtle and require close attention and sensitivity. For example, participants may be asked questions about their personal views, weaknesses, or fears. Such queries might lead people to reveal sensitive personal information. The point is not that researchers should refrain from asking questions, but that they need to be aware of the nature of the intrusion on people's psyches.

The need for sensitivity may be greater in qualitative studies, which often involve in-depth exploration on highly personal topics. In-depth probing may actually expose deep-seated fears that study participants had previously repressed. Qualitative researchers, regardless of the underlying research tradition, must be especially vigilant in anticipating such problems.

Example of intense self-scrutiny in a qualitative study: Caelli (2001) conducted a phenomenological study to illuminate nurses' understandings of health, and how such understandings translated into nursing practice. One participant, having explored her experience of health with the researcher over several interview sessions, resigned from her city hospital job as a result of gaining a new recognition of the role health played in her life.

The Right to Protection from Exploitation

Involvement in a study should not place participants at a disadvantage or expose them to damages. Participants need to be assured that their participation, or information they might provide, will not be used against them. For example, people describing their finances to a researcher should not be exposed to the risk of losing public healthcare benefits; those divulging illegal drug use should not fear exposure to criminal authorities.

Study participants enter into a special relationship with researchers, and it is crucial that this relationship not be exploited. Exploitation may be overt and malicious (e.g., sexual exploitation, use of donated blood for developing a commercial product), but might also be more subtle. For example, suppose people agreed to participate in a study requiring 30 minutes of their time and then the researcher decided 1 year later to go back to them, to follow their progress. Unless the researcher had previously warned participants that there might be a follow-up study, the researcher might be accused of not adhering to the agreement previously reached and of exploiting the researcher–participant relationship.

Because nurse researchers may have a nurse–patient (in addition to a researcher–participant) relationship, special care may be required to avoid exploiting that bond. Patients' consent to participate in a study may result from their understanding of the researcher's role as *nurse*, not as *researcher.*

In qualitative research, psychological distance between researchers and participants often declines as the study progresses. The emergence of a pseudotherapeutic relationship is not uncommon, which heightens the risk that exploitation could inadvertently occur (Eide & Kahn, 2008). On the other hand, qualitative researchers often are in a better position than quantitative researchers to *do good*, rather than just to avoid doing harm, because of the relationships they often develop with participants. Munhall (2012) has argued that qualitative nurse researchers have the responsibility of ensuring that, if there are any conflicts, the clinical and therapeutic imperative of nursing takes precedence over the research imperative of advancing knowledge.

Example of therapeutic research experiences: Beck (2005) reported that participants in her studies on birth trauma and post-traumatic stress disorder (PTSD) expressed a range of benefits from their e-mail exchanges with Beck. Here is what one informant voluntarily shared:

"You thanked me for everything in your e-mail, and I want to THANK YOU for caring. For me, it means a lot that you have taken an interest in this subject and are taking the time and effort to find out more about PTSD. For someone to even acknowledge this condition means a lot for someone who has suffered from it" (p. 417).

Respect for Human Dignity

Respect for human dignity is the second ethical principle in the *Belmont Report*. This principle includes the right to self-determination and the right to full disclosure.

The Right to Self-Determination

Humans should be treated as autonomous agents, capable of controlling their actions. **Self-determination** means that prospective participants can voluntarily decide whether to take part in a study, without risk of prejudicial treatment. It also means that people have the right to ask questions, to refuse to give information, and to withdraw from the study.

A person's right to self-determination includes freedom from **coercion**, which involves threats of penalty from failing to participate in a study or excessive rewards from agreeing to participate. Protecting people from coercion requires careful thought when the researcher is in a position of authority or influence over potential participants, as is often the case in a nurse–patient relationship. The issue of coercion may require scrutiny even when there is not a preestablished relationship. For example, a generous monetary incentive (or **stipend**) offered to encourage participation among an economically disadvantaged group (e.g., the homeless) might be considered mildly coercive because such incentives might pressure prospective participants into cooperation.

The Right to Full Disclosure

People's right to make informed, voluntary decisions about study participation requires full disclosure. **Full disclosure** means that the researcher has fully described the nature of the study, the person's right to refuse participation, the researcher's responsibilities, and likely risks and benefits. The right to self-determination and the right to full disclosure are the two major elements on which informed consent—discussed later in this chapter—is based.

Full disclosure is not always straightforward because it can create biases and sample recruitment problems. Suppose we were testing the hypothesis that high school students with a high rate of absenteeism are more likely to be substance abusers than students with good attendance. If we approached potential participants and fully explained the study purpose, some students likely would refuse to participate, and nonparticipation would be selective; those least likely to volunteer might well be substance abusing students—the group of primary interest. Moreover, by knowing the research question, those who *do* participate might not give candid responses. In such a situation, full disclosure could undermine the study.

A technique that is sometimes used in such situations is **covert data collection** (*concealment*), which is the collection of data without participants' knowledge and consent. This might happen, for example, if a researcher wanted to observe people's behavior in real-world settings and worried that doing so openly would affect the behavior of interest. Researchers might choose to obtain the information through concealed methods, such as by videotaping with hidden equipment or observing while pretending to be engaged in other activities. Covert data collection may in some cases be acceptable if risks are negligible and participants' right to privacy has not been violated. Covert data collection is least likely to be ethically tolerable if the study is focused on sensitive aspects of people's behavior, such as drug use or sexual conduct.

A more controversial technique is the use of **deception**, which involves deliberately withholding information about the study or providing participants with false information. For example, in studying high school students' use of drugs, we might describe the research as a study of students' health practices, which is a mild form of misinformation.

Deception and concealment are problematic ethically because they interfere with participants' right to make truly informed decisions about personal costs and benefits of participation. Some people argue that deception is never justified. Others, however, believe that if the study involves minimal risk to participants and if there are anticipated benefits to society, then deception may be justified to enhance the validity of the findings. ANA guidelines offer this advice about deception and concealment:

• • •

> The investigator understands that concealment or deception in research is controversial, depending on the type of research. Some investigators believe that concealment or deception in research can never be morally justified. The investigator further understands that before concealment or deception is used, certain criteria must be met: (1) The study must be of such small risk to the research participant and of such great significance to the advancement of the public good that concealment or deception can be morally justified . . . (2) The acceptability of concealment or deception is related to the degree of risks to research participants . . . (3) Concealment or deception are used only as last resorts, when no other approach can ensure the validity of the study's findings . . . (4) The investigator has a moral responsibility to inform research participants of any concealment or deception as soon as possible and to explain the rationale for its use. (Silva, 1995, p. 10, Section 4.2).

• • •

Another issue that has emerged in this era of electronic communication concerns data collection over the Internet. For example, some researchers analyze the content of messages posted to chat rooms, blogs, or listserves. The issue is whether such messages can be treated as research data without permission and informed consent. Some researchers believe that messages posted electronically are in the public domain and can be used without consent for research purposes. Others, however, feel that standard ethical rules should apply in cyberspace research and that electronic researchers must carefully protect the rights of those who are participants in "virtual" communities. Guidance for the ethical conduct of health research on the Internet has been developed by such writers as Ellett and colleagues (2004), Flicker and colleagues (2004), and Holmes (2009).

Justice

The third broad principle articulated in the *Belmont Report* concerns justice, which includes participants' right to fair treatment and their right to privacy.

The Right to Fair Treatment

One aspect of justice concerns the equitable distribution of benefits and burdens of research. Participant selection should be based on study requirements and not on a group's vulnerability. Participant selection has been a key ethical issue historically, with some researchers selecting groups with lower social standing (e.g., poor people, prisoners) as participants. The principle of justice imposes particular obligations toward individuals who are unable to protect their own interests (e.g., dying patients) to ensure that they are not exploited.

Distributive justice also imposes duties to neither neglect nor discriminate against individuals or groups who may benefit from research. During the 1980s and early 1990s, there was strong evidence that women and minorities were being unfairly excluded from many clinical studies in the United States. This led to the promulgation of regulations requiring that researchers who seek funding from the National Institutes of Health (NIH) include women and minorities as participants. The regulations also require researchers to examine whether clinical interventions have differential effects (e.g., whether benefits are different for men than for women), although this provision has had limited adherence (Polit & Beck, 2009).

The fair treatment principle covers issues other than participant selection. The right to fair treatment means that researchers must treat people who decline to participate (or who withdraw from the study after initial agreement) in a nonprejudicial manner; that they must honor all agreements made with participants (including payment of any promised

stipends); that they demonstrate respect for the beliefs, habits, and lifestyles of people from different backgrounds or cultures; that they give participants access to research staff for desired clarification; and that they afford participants courteous and tactful treatment at all times.

The Right to Privacy

Most research with humans involves intrusions into personal lives. Researchers should ensure that their research is not more intrusive than it needs to be and that participants' privacy is maintained continuously. Participants have the right to expect that their data will be kept in strictest confidence.

Privacy issues have become especially salient in the U.S. healthcare community since the passage of the Health Insurance Portability and Accountability Act of 1996 (HIPAA), which articulates federal standards to protect patients' health information. In response to the HIPAA legislation, the U.S. Department of Health and Human Services issued the regulations *Standards for Privacy of Individually Identifiable Health Information*. For most healthcare providers who transmit health information electronically, compliance with these regulations, known as the Privacy Rule, was required as of April 14, 2003.

TIP: Some information relevant to HIPAA compliance is presented in this chapter, but you should confer with any organizations that are involved in the research (if they are covered entities) regarding their practices and policies relating to HIPAA provisions. Also, there are websites that provide extensive information about the implications of HIPAA for health research: *http://privacyruleandresearch.nih.gov/* and *www.hhs.gov/ocr/hipaa/guidelines/research.pdf*.

PROCEDURES FOR PROTECTING STUDY PARTICIPANTS

Now that you are familiar with fundamental ethical principles in research, you need to understand procedures that researchers use to adhere to them.

Risk/Benefit Assessments

One strategy that researchers can use to protect participants is to conduct a **risk-benefit assessment**. Such an assessment is designed to examine whether the benefits of participating in a study are in line with the costs, be they financial, physical, emotional, or social—that is, whether the *risk/benefit* ratio is acceptable. The assessment of risks and benefits that individual participants might experience should be shared with them so that they can evaluate whether it is in their best interest to participate. Box 7.1 summarizes major costs and benefits of research participation.

TIP: The Toolkit in the accompanying *Resource Manual* includes a Word document with the factors in Box 7.1 arranged in worksheet form for you to complete in doing a risk/benefit assessment. By completing the worksheet, it may be easier for you to envision opportunities for "doing good" and to avoid possibilities of doing harm.

The risk/benefit ratio should also consider whether risks to participants are on a par with benefits to society and to nursing in terms of the evidence produced. A broad guideline is that the degree of risk by participants should never exceed the potential humanitarian benefits of the knowledge to be gained. Thus, the selection of a significant topic that has the potential to improve patient care is the first step in ensuring that research is ethical.

All research involves some risks, but risk is sometimes minimal. **Minimal risk** is defined as risks no greater than those ordinarily encountered in daily life or during routine tests or procedures. When the risks are not minimal, researchers must proceed with caution, taking every step possible to diminish risks and maximize benefits. If expected risks to participants outweigh the anticipated benefits of the study, the research should be redesigned.

In quantitative studies, most details of the study usually are spelled out in advance, so a reasonably accurate risk/benefit ratio assessment can be developed. Qualitative studies, however, usually evolve as data are gathered, so it may be more difficult to

BOX 7.1 Potential Benefits and Risks of Research to Participants

MAJOR POTENTIAL BENEFITS TO PARTICIPANTS

- Access to a potentially beneficial intervention that might otherwise be unavailable to them
- Comfort in being able to discuss their situation or problem with a friendly, objective person
- Increased knowledge about themselves or their conditions, either through opportunity for introspection and self-reflection or through direct interaction with researchers
- Escape from normal routine, excitement of being part of a study
- Satisfaction that information they provide may help others with similar problems or conditions
- Direct monetary or material gains through stipends or other incentives

MAJOR POTENTIAL RISKS TO PARTICIPANTS

- Physical harm, including unanticipated side effects
- Physical discomfort, fatigue, or boredom
- Psychological or emotional distress resulting from self-disclosure, introspection, fear of the unknown, discomfort with strangers, fear of eventual repercussions, anger or embarrassment at the type of questions being asked
- Social risks, such as the risk of stigma, adverse effects on personal relationships, loss of status
- Loss of privacy
- Loss of time
- Monetary costs (e.g., for transportation, child care, time lost from work)

assess all risks at the outset. Qualitative researchers must remain sensitive to potential risks throughout the study.

Example of ongoing risk/benefit assessment: Carlsson and colleagues (2007) discussed ethical issues relating to the conduct of interviews with people who have brain damage. The researchers noted the need for ongoing vigilance and attention to cues about risks and benefits. For example, one interview had to be interrupted because the participant displayed signs of distress. Afterward, however, the participant expressed gratitude for the opportunity to discuss his experience.

One potential benefit to participants is monetary. Stipends offered to prospective participants are rarely viewed as an opportunity for financial gain, but there is ample evidence that stipends are useful incentives to participant recruitment and retention (Edwards et al., 2009; Robinson et al., 2007). Financial incentives are especially effective when the group under study is difficult to recruit, when the study is time-consuming or tedious, or when participants incur study-related costs (e.g., for child care or transportation). Stipends range from $1 to hundreds of dollars, but most are in the $20 to $30 range.

TIP: In evaluating the anticipated risk/benefit ratio of a study design, you might want to consider how comfortable *you* would feel about being a study participant.

Informed Consent and Participant Authorization

A particularly important procedure for safeguarding study participants involves obtaining their informed consent. **Informed consent** means that participants have adequate information about the research, comprehend that information, and have the ability to consent to or decline participation voluntarily. This section discusses procedures for

obtaining informed consent and for complying with HIPAA rules regarding accessing patients' health information.

The Content of Informed Consent

Fully informed consent involves communicating the following pieces of information to participants:

1. *Participant status.* Prospective participants need to understand the distinction between *research* and *treatment.* They should be told which healthcare activities are routine and which are implemented specifically for the study. They also should be informed that data they provide will be used for research purposes.
2. *Study goals.* The overall goals of the research should be stated, in lay rather than technical terms. The use to which the data will be put should be described.
3. *Type of data.* Prospective participants should be told what type of data will be collected.
4. *Procedures.* Prospective participants should be given a description of the data collection procedures and of procedures to be used in any innovative treatment.
5. *Nature of the commitment.* Participants should be told the expected time commitment at each point of contact and the number of contacts within a given timeframe.
6. *Sponsorship.* Information on who is sponsoring or funding the study should be noted; if the research is part of an academic requirement, this information should be shared.
7. *Participant selection.* Prospective participants should be told how they were selected for recruitment and how many people will be participating.
8. *Potential risks.* Prospective participants should be informed of any foreseeable risks (physical, psychological, social, or economic) or discomforts and efforts that will be taken to minimize risks. The possibility of unforeseeable risks should also be discussed, if appropriate. If injury or damage is possible, treatments that will be made available to participants should be described. When risks are more than minimal, prospective participants should be encouraged to seek advice before consenting.
9. *Potential benefits.* Specific benefits to participants, if any, should be described, as well as possible benefits to others.
10. *Alternatives.* If appropriate, participants should be told about alternative procedures or treatments that might be advantageous to them.
11. *Compensation.* If stipends or reimbursements are to be paid (or if treatments are offered without fee), these arrangements should be discussed.
12. *Confidentiality pledge.* Prospective participants should be assured that their privacy will at all times be protected. If anonymity can be guaranteed, this should be stated.
13. *Voluntary consent.* Researchers should indicate that participation is strictly voluntary and that failure to volunteer will not result in any penalty or loss of benefits.
14. *Right to withdraw and withhold information.* Prospective participants should be told that, after consenting, they have the right to withdraw from the study or to withhold any specific piece of information. Researchers may need to describe circumstances under which researchers would terminate the study.
15. *Contact information.* The researcher should tell participants whom they could contact in the event of further questions, comments, or complaints.

In qualitative studies, especially those requiring repeated contact with participants, it may be difficult to obtain meaningful informed consent at the outset. Qualitative researchers do not always know in advance how the study will evolve. Because the research design emerges during data collection, researchers may not know the exact nature of the data to be collected, what the risks and benefits to participants will be, or how much of a time commitment they will be expected to make. Thus, in a qualitative study, consent is often viewed as an ongoing, transactional process, sometimes called **process consent**. In process consent, the researcher continually renegotiates the consent, allowing participants

to play a collaborative role in the decision-making process regarding ongoing participation.

Example of process consent: Treacy and colleagues (2007) conducted a three-round longitudinal study of children's emerging perspectives and experiences of cigarette smoking. Parents and children consented to the children's participation. At each round, consent to continue participating in the study was reconfirmed.

Comprehension of Informed Consent

Consent information is normally presented to prospective participants while they are being recruited, either orally or in writing. Written notices should not, however, take the place of spoken explanations, which provide opportunities for elaboration and for participants to question and "screen" the researchers.

Example of "screening" of researchers: Speraw (2009) did an in-depth study of adults and children with disabilities. Parental consent was obtained for child participants, and Speraw noted that:

". . . extensive discussion with parents took place via telephone. Additional conversations took place in the participants' homes prior to the interview. This period of rapport building was deemed essential, allowing parents ample opportunity to screen the researcher and make a determination of the suitability of the study for their child" (p. 736).

Because informed consent is based on a person's evaluation of the potential risks and benefits of participation, critical information must not only be communicated, but also understood. Researchers may have to play a "teacher" role in communicating consent information. They should be careful to use simple language and to avoid jargon and technical terms whenever possible; they should also avoid language that might unduly influence the person's decision to participate. Written statements should be consistent with the participants' reading levels and educational attainment. For participants from a general population (e.g., patients in a hospital), the statement should be written at about the 7th or 8th grade reading level.

TIP: Yates and colleagues (2009) described an innovative visual presentation of informed consent information designed to improve communication and enhance participation rates.

For some studies, especially those involving more than minimal risk, researchers need to make special efforts to ensure that prospective participants understand what participation will entail. In some cases, this might involve testing participants for their comprehension of the informed consent material before deeming them eligible. Such efforts are especially warranted with participants whose native tongue is not English or who have cognitive impairments.

Example of confirming comprehension in informed consent: Horgas and colleagues (2008) studied the relationship between pain and functional disability in older adults. Prospective participants had to demonstrate ability to provide informed consent:

"Ability to consent was ascertained by explaining the study to potential participants, who were then asked to describe the study" (p. 344). All written materials for the study, including consent forms, were at the 8th-grade reading level and printed in 14-point font.

Documentation of Informed Consent

Researchers usually document informed consent by having participants sign a **consent form**. In the United States, federal regulations for studies funded by the government require written consent of participants, except under certain circumstances. When the study does not involve an intervention and data are collected anonymously—or when existing data from records or specimens are used and identifying information is not linked to the data—regulations requiring written informed consent do not apply. HIPAA legislation is explicit about the type of information that must be eliminated from patient records for the data to be considered **de-identified**.

The consent form should contain all the information essential to informed consent. Prospective participants (or a legally authorized representative) should have ample time to review the document

before signing it. The consent form should also be signed by the researcher, and a copy should be retained by both parties.

An example of a written consent form used in a study of one of the authors is presented in Figure 7.1. The numbers in the margins of this figure correspond to the types of information for informed consent outlined earlier. (The form does not indicate how people were selected; prospective participants knew they were recruited from a particular support group.)

TIP: In developing a consent form, the following suggestions might prove helpful:

1. Organize the form coherently so that prospective participants can follow the logic of what is being communicated. If the form is complex, use headings as an organizational aid.
2. Use a large enough font so that the form can be easily read, and use spacing that avoids making the document appear too dense. Make the form attractive and inviting.
3. In general, simplify. Use clear, consistent terminology. Avoid technical terms if possible. If technical terms are needed, include definitions. Some suggestions are offered in the Toolkit.
4. Assess the form's reading level by using a **readability formula** to ensure an appropriate level for the group under study. There are several such formulas, the most widely used being the FOG Index (Gunning, 1968), the Flesch Reading Ease score, and Flesch-Kincaid grade level score (Flesch, 1948). Microsoft Word provides Flesch readability statistics.
   - In Word 2003, click Tools → Options → Spelling and Grammar → Show Readability Statistics.
   - In Word 2007, click the Microsoft Office button (upper left corner) → Word Options → Proofing → Check Grammar with Spelling + Show Readability Statistics.
   - In Word 2010, click the blue Office button (upper left corner) → Options → Proofing → Check Grammar with Spelling + Show Readability Statistics.
5. Test the form with people similar to those who will be recruited, and ask for feedback.

In certain circumstances (e.g., with non–English-speaking participants), researchers with NIH funding have the option of presenting the full information orally and then summarizing essential information in a **short form**. If a short form is used, however, the oral presentation must be witnessed by a third party, and the witness's signature must appear on the short consent form. The signature of a third-party witness is also advisable in studies involving more than minimal risk, even when a comprehensive consent form is used.

When the primary means of data collection is through a self-administered questionnaire, some researchers do not obtain written informed consent because they assume **implied consent** (i.e., that the return of the completed questionnaire reflects voluntary consent to participate). This assumption, however, may not always be warranted (e.g., if patients feel that their treatment might be affected by failure to cooperate with the researcher).

TIP: The Toolkit in the accompanying *Resource Manual* includes several informed consent forms as Word documents that can be adapted for your use. (Many universities offer templates for consent forms.) The Toolkit also includes several other resources designed to help you with the ethical aspects of a study.

Authorization to Access Private Health Information

Under HIPAA regulations in the United States, a covered entity such as a hospital can disclose individually identifiable health information (IIHI) from its records if the patient signs an authorization. The authorization can be incorporated into the consent form, or it can be a separate document. Using a separate authorization form may be advantageous to protect the patients' confidentiality because the form does not need to provide detailed information about the purpose of the research. If the research purpose is not sensitive, or if the hospital or entity is already cognizant of the study purpose, an integrated authorization and consent form may suffice.

The authorization, whether obtained separately or as part of the consent form, must include the following: (1) who will receive the information, (2) what type of information will be disclosed, and (3) what further disclosures the researcher

Informed Consent Form

1
2
3,5
4
12
11
8

I understand that I am being asked to participate in a research study at Saint Francis Hospital and Medical Center. This research study will evaluate: What it is like being a mother of multiples during the first year of the infants' lives. If I agree to participate in the study, I will be interviewed for approximately 30 to 60 minutes about my experience as a mother of multiple infants. The interview will be tape-recorded and take place in a private office at Saint Francis Hospital. No identifying information will be included when the interview is transcribed. I understand I will receive $25.00 for participating in the study. There are no known risks associated with this study.

7 I realize that I may not participate in the study if I am younger than 18 years of age or I cannot speak English.

10 I realize that the knowledge gained from this study may help either me or other mothers of multiple infants in the future.

13
14

I realize that my participation in this study is entirely voluntary, and I may withdraw from the study at any time I wish. If I decide to discontinue my participation in this study, I will continue to be treated in the usual and customary fashion.

12 I understand that all study data will be kept confidential. However, this information may be used in nursing publications or presentations.

8 I understand that if I sustain injuries from my participation in this research project, I will not be automatically compensated by Saint Francis Hospital and Medical Center.

15 If I need to, I can contact Dr. Cheryl Beck, University of Connecticut, School of Nursing, any time during the study.

1,2 The study has been explained to me. I have read and understand this consent form, all of my questions have been answered, and I agree to participate. I understand that I will be given a copy of this signed consent form.

_____ _____

Signature of Participant Date

_____ _____

Signature of Witness Date

_____ _____

Signature of Investigator Date

FIGURE 7.1 Example of an informed consent form.

anticipates. The need for patient authorization to access IIHI can be waived only under certain circumstances. Patient authorization usually must be obtained for data that are *created* as part of the research, as well as for information already maintained in institutional records (Olsen, 2003).

Confidentiality Procedures

Study participants have the right to expect that data they provide will be kept in strict confidence. Participants' right to privacy is protected through various confidentiality procedures.

Anonymity

Anonymity, the most secure means of protecting confidentiality, occurs when the researcher cannot link participants to their data. For example, if questionnaires were distributed to a group of nursing home residents and were returned without any identifying information, responses would be anonymous. As another example, if a researcher reviewed hospital records from which all identifying information (e.g., name, social security number, and so on) had been expunged, anonymity would again protect participants' right to privacy. Whenever it is possible to achieve anonymity, researchers should strive to do so. Distributed questionnaires through the mail, to groups of participants, or over the Internet are especially conducive to anonymity.

Example of anonymity: Wagner and colleagues (2009) distributed anonymous questionnaires to members of gerontological nursing organizations in the United States and Canada. The questionnaires elicited nurses' perceptions of workplace safety culture in long-term care settings.

Confidentiality in the Absence of Anonymity

When anonymity is impossible, confidentiality procedures need to be implemented. A promise of **confidentiality** is a pledge that any information participants provide will not be publicly reported in a manner that identifies them, and will not be accessible to others. This means that research information should not be shared with strangers nor with people known to participants (e.g., relatives, doctors, other nurses), unless participants give explicit permission to do so.

Researchers can take a number of steps to ensure that a *breach of confidentiality* does not occur, including the following:

- Obtain identifying information (e.g., name, address) from participants only when essential.
- Assign an **identification** (ID) **number** to each participant and attach the ID number rather than other identifiers to the actual data.
- Maintain identifying information in a locked file.
- Restrict access to identifying information to only a few people on a need-to-know basis.
- Enter no identifying information onto computer files.
- Destroy identifying information as quickly as practical.
- Make research personnel sign confidentiality pledges if they have access to data or identifying information.
- Report research information in the aggregate; if information for an individual is reported, disguise the person's identity, such as through the use of a fictitious name.

TIP: Researchers who plan to collect data from participants multiple times (or who use multiple forms that need to be linked) do not have to forego anonymity. A technique that has been successful is to have participants themselves generate an ID number. They might be instructed, for example, to use their birth year and the first three letters of their mother's maiden names as their ID code (e.g., 1946CRU). This code would be put on every form so that forms could be linked, but researchers would not know participants' identities.

Qualitative researchers may need to take extra steps to safeguard participants' privacy. Anonymity is almost never possible in qualitative studies because researchers typically become closely involved with participants. Moreover, because of the in-depth nature of qualitative studies, there may be a greater invasion of privacy than is true in quantitative research. Researchers who spend time in the home

of a participant may, for example, have difficulty segregating the public behaviors that the participant is willing to share from private behaviors that unfold during data collection. A final issue is adequately disguising participants in reports. Because the number of participants is small, qualitative researchers may need to take extra precautions to safeguard identities. This may mean more than simply using a fictitious name. Qualitative researchers may have to slightly distort identifying information, or provide only general descriptions. For example, a 49-year-old antique dealer with ovarian cancer might be described as "a middle-aged cancer patient who worked in retail sales" to avoid identification that could occur with the more detailed description.

Example of confidentiality procedures in a qualitative study: Graffigna and Olson (2009) studied how young people talk about HIV/AIDS in a group interview. Potential participants were assured of confidentiality and the voluntary nature of participation. Participants signed consent forms in the presence of researchers so that questions could be addressed. Names and identifying information were removed from data and stored separately in the researchers' office. Transcripts of the group discussion were analyzed anonymously.

Certificates of Confidentiality

There are situations in which confidentiality can create tensions between researchers and legal or other authorities, especially if participants are involved in criminal or dangerous activity (e.g., substance abuse, unprotected sexual intercourse). To avoid the possibility of forced, involuntary disclosure of sensitive research information (e.g., through a court order or subpoena), researchers in the United States can apply for a **Certificate of Confidentiality** from the National Institutes of Health (Lutz et al., 2000). Any research that involves the collection of personally identifiable, sensitive information is potentially eligible for a Certificate, even if the study is not federally funded. Information is considered sensitive if its release might damage participants' financial standing, employability, or reputation or might lead to discrimination; information about a person's mental health, as well as genetic information, is also considered sensitive.

A Certificate of Confidentiality protects against the forced disclosure of research data in a wide range of situations. A Certificate allows researchers to refuse to disclose identifying information on study participants in any civil, criminal, administrative, or legislative proceeding at the federal, state, or local level.

A Certificate of Confidentiality helps researchers to achieve their research objectives without threat of involuntary disclosure and can be helpful in recruiting participants. Researchers who obtain a Certificate should alert prospective participants about this valuable protection in the consent form, and should note any planned exceptions to those protections. For example, a researcher might decide to voluntarily comply with state child abuse reporting laws even though the Certificate would prevent authorities from punishing researchers who chose not to comply.

Example of obtaining a Certificate of Confidentiality: Laughon (2007) conducted an in-depth study of the ways in which poor, urban African American women with a history of physical abuse stay healthy. Interviews covered a range of sensitive topics (domestic violence, substance abuse), so the researcher obtained a Certificate of Confidentiality.

Debriefings, Communications, and Referrals

Researchers can often show their respect for participants—and proactively minimize emotional risks—by carefully attending to the nature of the interactions they have with them. For example, researchers should always be gracious and polite, should phrase questions tactfully, and should be sensitive to cultural and linguistic diversity.

Researchers can also use more formal strategies to communicate respect and concern for participants' well-being. For example, it is sometimes useful to offer **debriefing** sessions after data collection is completed to permit participants to ask questions or air complaints. Debriefing is especially important when the data collection has been stressful or when ethical guidelines had to be "bent" (e.g., if any deception was used in explaining the study).

Example of debriefing: Sandgren and colleagues (2006) studied strategies that palliative cancer nurses used to avoid being emotionally overloaded. After each in-depth interview with 46 nurses,

"... we made sure that the participants were doing well, and we assessed possible needs for emotional support" (p. 81).

It is also thoughtful to communicate with participants after the study is completed to let them know that their participation was appreciated. Researchers sometimes demonstrate their interest in study participants by offering to share study findings with them once the data have been analyzed (e.g., by mailing them a summary or advising them of an appropriate website).

Example of thanking participants: Hsiao and Van Riper (2009) studied individual and family adaptation in Taiwanese families with relatives who had severe and persistent mental illness. At the end of the study, each participant was sent a thank you card to convey gratitude for their time.

Finally, in some situations, researchers may need to assist study participants by making referrals to appropriate health, social, or psychological services.

Example of referrals: Caldwell and Redeker (2009) studied psychological distress in women living in inner cities. All participants were offered the opportunity to obtain counseling at a local health center. Women whose psychological distress scores were moderate were referred to the health center. Those whose scores were severe were escorted to the psychiatric emergency room where they were immediately evaluated by a clinician.

Treatment of Vulnerable Groups

Adherence to ethical standards is often straightforward, but additional procedures and heightened sensitivity may be required to protect the rights of special vulnerable groups. **Vulnerable populations** may be incapable of giving fully informed consent (e.g., mentally retarded people) or may be at risk of unintended side effects because of their circumstances (e.g., pregnant women). Researchers interested in studying high-risk groups should understand guidelines governing informed consent, risk/benefit assessments, and acceptable research procedures for such groups. In general, research with vulnerable groups should be undertaken only when the risk/benefit ratio is low or when there is no alternative (e.g., studies of childhood development require child participants).

Among the groups that nurse researchers should consider vulnerable are the following:

- *Children.* Legally and ethically, children do not have competence to give informed consent, so the informed consent of children's parents or legal guardians must be obtained. It is appropriate, however—especially if the child is at least 7 years old—to obtain the child's assent as well. **Assent** refers to the child's affirmative agreement to participate. If the child is mature enough (e.g., a 12-year-old) to understand basic informed consent information, it is advisable to obtain written assent from the child as well, as evidence of respect for the child's right to self-determination. Lindeke and colleagues (2000) and Kanner and colleagues (2004) provided guidance on children's assent and consent to participate in research. The U.S. government has issued special regulations (Subpart D of the Code of Federal Regulations, 2005) for the additional protection of children as study participants.
- *Mentally or emotionally disabled people.* Individuals whose disability makes it impossible for them to weigh the risks and benefits of participation (e.g., people affected by cognitive impairment, coma, and so on) also cannot legally or ethically provide informed consent. In such cases, researchers should obtain the written consent of a legal guardian. To the extent possible, informed consent or assent from participants themselves should be sought as a supplement to consent by a guardian. NIH guidelines note that studies involving people whose autonomy is compromised by disability should focus in a direct way on their condition.
- *Severely ill or physically disabled people.* For patients who are very ill or undergoing certain treatments, it might be necessary to assess their ability to make reasoned decisions about study participation. For example, Higgins and Daly

(1999) described a process they used to assess decisional capacity in mechanically ventilated patients. For certain disabilities, special procedures for obtaining consent may be required. For example, with deaf participants, the entire consent process may need to be in writing. For people who have a physical impairment preventing them from writing or for participants who cannot read and write, alternative procedures for documenting informed consent (such as audiotaping or videotaping consent proceedings) should be used.

- *The terminally ill.* Terminally ill people who participate in studies seldom expect to benefit personally from the research, so the risk/benefit ratio needs to be carefully assessed. Researchers must also take steps to ensure that the healthcare and comfort of terminally ill participants are not compromised. Special procedures may be needed to obtain informed consent if they are physically or mentally incapacitated.
- *Institutionalized people.* Particular care is required in recruiting institutionalized people because they depend on healthcare personnel and may feel pressured into participating, or may believe that their treatment would be jeopardized by failure to cooperate. Inmates of prisons and other correctional facilities, who have lost their autonomy in many spheres of activity, may similarly feel constrained in their ability to withhold consent. The U.S. government has issued specific regulations for the protection of prisoners as study participants (see Code of Federal Regulations, 2005, Subpart C). Researchers studying institutionalized groups need to emphasize the voluntary nature of participation.
- *Pregnant women.* The U.S. government has issued additional requirements governing research with pregnant women and fetuses (Code of Federal Regulations, 2005, Subpart B). These requirements reflect a desire to safeguard both the pregnant woman, who may be at heightened physical and psychological risk, and the fetus, who cannot give informed consent. The regulations stipulate that a pregnant woman cannot be involved in a study unless its purpose is to meet the health needs of the pregnant woman, and risks to her and the fetus are minimized or there is only a minimal risk to the fetus.

Example of research with a vulnerable group: Kelly and colleagues (2009) studied dating violence among girls (average age of 15) in the juvenile justice system who were participating in a health promotion program in Bexar County, Texas. The authors noted that because of the high prevalence of violence and neglect in this population, the ethics review committee of Kelly's university waived obtaining parental consent as being a source of potential harm. Girls were assured in person that participation was voluntary and that lack of participation would not affect their detention or probation status.

It should go without saying that researchers need to proceed with great caution in conducting research with people who might fall into two or more vulnerable categories, as was the case in the preceding example.

TIP: Jacobson (2005) has astutely pointed out the need to be vigilant on behalf of persons not traditionally identified as vulnerable and, therefore, not covered in standard protocols regarding vulnerable participants. *Anybody* may be vulnerable at any given time due to acute illness or special circumstances that challenge the capacity to provide truly informed consent.

External Reviews and the Protection of Human Rights

Researchers, who often have a strong commitment to their research, may not be objective in their risk/benefit assessments or in their efforts to protect participants' rights. Because of the possibility of a biased self-evaluation, the ethical dimensions of a study should normally be subjected to external review.

Most institutions where research is conducted have formal committees for reviewing proposed research plans. These committees are sometimes called *human subjects committees*, *ethical advisory boards*, or *research ethics committees.* In the United States, the committee likely will be called an **Institutional Review Board (IRB)**, whereas in Canada it is called a **Research Ethics Board (REB)**.

TIP: You should find out early what an institution's requirements are regarding ethics, in terms of its forms, procedures, and review schedules. Also, it is wise to allow a generous amount of time for negotiating with IRBs, which may require procedural modifications and re-review.

Qualitative researchers in various countries have expressed some concerns that standard ethical review procedures are not sensitive to special issues and circumstances faced in qualitative research. There is concern that regulations were " . . . created for quantitative work, and can actually impede or interrupt work that is not hypothesis-driven 'hard science'" (Van de Hoonaard, 2002, p. i). Thus, qualitative researchers may need to take extra care to explain their methods, rationales, and approaches to review board members unfamiliar with qualitative research.

Institutional Review Boards

In the United States, federally sponsored studies are subject to strict guidelines for evaluating the treatment of human participants. (Guidance on human subjects issues in grant applications is provided in Chapter 29.) Before undertaking such a study, researchers must submit research plans to the IRB, and must also go through formal training on ethical conduct and a certification process that can be completed online.

The duty of the IRB is to ensure that the proposed plans meet federal requirements for ethical research. An IRB can approve the proposed plans, require modifications, or disapprove the plans. The main requirements governing IRB decisions may be summarized as follows (Code of Federal Regulations, 2005, §46.111):

- Risks to participants are minimized.
- Risks to participants are reasonable in relation to anticipated benefits, if any, and the importance of the knowledge that may reasonably be expected to result.
- Selection of participants is equitable.
- Informed consent will be sought, as required, and appropriately documented.
- Adequate provision is made for monitoring the research to ensure participants' safety.
- Appropriate provisions are made to protect participants' privacy and confidentiality of the data.
- When vulnerable groups are involved, appropriate additional safeguards are included to protect their rights and welfare.

Example of IRB approval: Jones and her colleagues (2010) studied the meaning of surviving cancer among Latino adolescents and young adults. The procedures and protocols for the study were approved by the IRBs of two cancer clinics where the study was conducted.

Many studies require a full IRB review involving a meeting at which a majority of IRB members are present. An IRB must have five or more members, at least one of whom is not a researcher (e.g., a member of the clergy or a lawyer may be appropriate). One IRB member must be a person who is not affiliated with the institution and is not a family member of an affiliated person. To protect against potential biases, the IRB cannot comprise entirely men, women, or members from a single profession.

For certain research involving no more than minimal risk, the IRB can use expedited review procedures, which do not require a meeting. In an **expedited review**, a single IRB member (usually the IRB chairperson) carries out the review. An example of research that qualifies for an expedited IRB review is minimal-risk research ". . . employing survey, interview, focus group, program evaluation, human factors evaluation, or quality assurance methodologies" (Code of Federal Regulations, 2005, §46.110).

Federal regulations also allow certain types of research in which there are no apparent risk to participants to be exempt from IRB review. The website of the Office of Human Research Protections, in its policy guidance section, includes decision charts designed to clarify whether a study is exempt.

TIP: Researchers seeking a Certificate of Confidentiality must first obtain IRB approval because such approval is a prerequisite for the Certificate. Applications for the Certificate should be submitted at least 3 months before participants are expected to enroll in the study.

Data and Safety Monitoring Boards

In addition to IRBs, researchers in the United States may have to communicate information about ethical aspects of their studies to other groups. For example, some institutions have established separate **Privacy Boards** to review researchers' compliance with provisions in HIPAA, including review of authorization forms and requests for waivers.

For researchers evaluating interventions in clinical trials, NIH also requires review by a **data and safety monitoring board** (DSMB). The purpose of a DSMB is to oversee the safety of participants, to promote data integrity, and to review accumulated outcome data on a regular basis to determine whether study protocols should be altered, or the study stopped altogether. Members of a DSMB are selected based on their clinical, statistical, and methodologic expertise. The degree of monitoring by the DSMB should be proportionate to the degree of risk involved.

Example of a Data and Safety Monitoring Board: Artinian and colleagues (2007) tested the effectiveness of a nurse-managed telemonitoring intervention for lowering blood pressure among hypertensive African Americans. In a separate article, the researchers presented a good description of their data and safety monitoring plan and discussed how IRBs and DSMBs differ (Artinian et al., 2004).

Building Ethics into the Design of the Study

Researchers need to give careful thought to ethical requirements while planning a study, and should ask themselves whether intended safeguards for protecting humans are sufficient. They must continue their vigilance throughout the course of the study as well, because unforeseen ethical dilemmas may arise. Of course, first steps in doing ethical research include scrutinizing the research question to determine if it is clinically significant and designing the study in a manner that yields sound evidence—it can be construed as unethical to do poorly conceived or weakly designed research because it would be a poor use of people's time.

The remaining chapters of the book offer advice on how to design studies that yield high-quality evidence for practice. Methodologic decisions about rigor, however, must be made within the context of ethical requirements. Box 7.2 presents some examples of the kinds of questions that might be posed in thinking about ethical aspects of study design.

TIP: After study procedures have been developed, researchers should undertake a self-evaluation of those procedures to determine if they meet ethical requirements. Box 7.3, later in this chapter, provides some guidelines that can be used for such a self-evaluation.

OTHER ETHICAL ISSUES

In discussing ethical issues relating to the conduct of nursing research, we have given primary consideration to the protection of human participants. Two other ethical issues also deserve mention: the treatment of animals in research and research misconduct.

Ethical Issues in Using Animals in Research

Some nurse researchers use animals rather than human beings as their subjects, typically focusing on biophysiologic phenomena. Despite some opposition to such research by animal rights activists, researchers in health fields likely will continue to use animals to explore physiologic mechanisms and to test interventions that could pose risks to humans.

Ethical considerations are clearly different for animals and humans; for example, the concept of *informed consent* is not relevant for animal subjects. Guidelines have been developed governing treatment of animals in research. In the United States, the Public Health Service issued a policy statement on the humane care and use of animals, most recently amended in 2002. The guidelines articulate nine principles for the proper treatment of animals used in

BOX 7.2 Examples of Questions for Building Ethics into a Study Design

RESEARCH DESIGN

- Will participants get allocated fairly to different treatment groups?
- Will steps to reduce bias or enhance integrity add to the risks participants will incur?
- Will the setting for the study protect against participant discomfort?

INTERVENTION

- Is the intervention designed to maximize good and minimize harm?
- Under what conditions might a treatment be withdrawn or altered?

SAMPLE

- Is the population defined so as to unwittingly and unnecessarily exclude important segments of people (e.g., women or minorities)?
- Will potential participants be recruited into the study equitably?

DATA COLLECTION

- Will data be collected in such a way as to minimize respondent burden?
- Will procedures for ensuring confidentiality of data be adequate?
- Will data collection staff be appropriately trained to be sensitive and courteous?

REPORTING

- Will participants' identities be adequately protected?

biomedical and behavioral research. These principles cover such issues as the transport of research animals, alternatives to using animals, pain and distress in animal subjects, researcher qualifications, the use of appropriate anesthesia, and euthanizing animals under certain conditions. In Canada, researchers who use animals in their studies must adhere to the policies and guidelines of the Canadian Council on Animal Care (CCAC) as articulated in the two-volume *Guide to the Care and Use of Experimental Animals.*

Holtzclaw and Hanneman (2002) noted several important considerations in the use of animals in nursing research. First, there must be a compelling reason to use an animal model—not simply convenience or novelty. Second, study procedures should be humane, well planned, and well funded. Animal studies are not necessarily less costly than those with human participants, and they require serious ethical and scientific consideration to justify their use.

Example of research with animals: Raines and other nurse anesthetists (2009) studied the anxiolytic effects of luteolin, a lemon balm flavenoid, in male Sprague-Dawley rats. In all, 55 rats were used in the study. Protocols for the use of the rats were in accordance with NIH's Guide for the Care and Use of Laboratory Animals and they received approval from an Institutional Animal Care and Use Committee.

Research Misconduct

Ethics in research involves not only the protection of human and animal subjects, but also protection of the public trust. The issue of **research misconduct** (or *scientific misconduct)* has received greater attention in recent years as incidents of researcher fraud and misrepresentation have come to light. Currently, the U.S. agency responsible for overseeing efforts to improve research integrity and for handling

allegations of research misconduct is the Office of Research Integrity (ORI) within DHHS. Researchers seeking funding from NIH must demonstrate that they have received training on research integrity and the responsible conduct of research.

Research misconduct, as defined by a 2005 U.S. Public Health Service regulation (42 CFR Part 93), is "fabrication, falsification, or plagiarism in proposing, performing, or reviewing research, or in reporting research results." To be construed as misconduct, there must be a significant departure from accepted practices in the research community, and the misconduct must have been committed intentionally, knowingly, or recklessly. *Fabrication* involves making up data or study results. *Falsification* involves manipulating research materials, equipment, or processes; it also involves changing or omitting data, or distorting results such that the research is not accurately represented in reports. *Plagiarism* involves the appropriation of someone's ideas, results, or words without giving due credit, including information obtained through the confidential review of research proposals or manuscripts.

Although the official definition focuses on only three types of misconduct, there is widespread agreement that research misconduct covers many other issues including improprieties of authorship, poor data management, conflicts of interest, inappropriate financial arrangements, failure to comply with governmental regulations, and unauthorized use of confidential information. Conflicts of interest may be a particularly salient issue in health-related research funded by for-profit organizations.

Example of research misconduct: In 2008, the U.S. Office of Research Integrity ruled that a nurse in Missouri engaged in scientific misconduct in research supported by the National Cancer Institute. The nurse falsified and fabricated data that were reported to the National Surgical Adjuvant Breast and Bowel Project (NIH Notice Number NOT-OD-08-096).

Research integrity is an important concern in nursing. Jeffers and Whittemore (2005), for example, engaged in work to identify and describe research environments that promote integrity. In a study that focused on ethical issues faced by editors of nursing journals, Freda and Kearney (2005) found that 64% of the 88 editors reported some type of ethical dilemma, such as duplicate publication, plagiarism, or conflicts of interest. Editors in several major nursing journals subsequently wrote editorials about this topic (e.g., Baggs, 2008; Broome, 2008). Habermann and colleagues (2010) studied 1,645 research coordinators' experiences with research misconduct in their clinical environments. More than 250 coordinators, most of them nurses, said they had first-hand knowledge of scientific misconduct that included protocol violations, consent violations, fabrication, falsification, and financial conflicts of interest.

Example of research on research integrity: In 2005, Gwen Anderson was awarded a grant through NINR under its Research on Research Integrity initiative. Her study explored common daily practices and systems in gene therapy clinical research, and sought to describe institutional cultures that promote or protect research integrity—as well as those that do not. In another study, Dr. Anderson (2008) examined the ethical preparedness and performance of gene therapy study coordinators.

CRITIQUING THE ETHICS OF RESEARCH STUDIES

Guidelines for critiquing ethical aspects of a study are presented in Box 7.3. Members of an ethics committee should be provided with sufficient information to answer all these questions. Research journal articles, however, do not always include detailed information about ethics because of space constraints. Thus, it is not always possible to critique researchers' adherence to ethical guidelines, but we offer a few suggestions for considering a study's ethical aspects.

Many research reports acknowledge that study procedures were reviewed by an IRB or ethics committee. When a report specifically mentions a formal review, it is usually safe to assume that a group of concerned people did a conscientious review of the study's ethical issues.

BOX 7.3 Guidelines for Critiquing the Ethical Aspects of a Study

1. Was the study approved and monitored by an Institutional Review Board, Research Ethics Board, or other similar ethics review committee?
2. Were participants subjected to any physical harm, discomfort, or psychological distress? Did the researchers take appropriate steps to remove, prevent, or minimize harm?
3. Did the benefits to participants outweigh any potential risks or actual discomfort they experienced? Did the benefits to society outweigh the costs to participants?
4. Was any type of coercion or undue influence used to recruit participants? Did they have the right to refuse to participate or to withdraw without penalty?
5. Were participants deceived in any way? Were they fully aware of participating in a study and did they understand the purpose and nature of the research?
6. Were appropriate informed consent procedures used? If not, were there valid and justifiable reasons?
7. Were adequate steps taken to safeguard participants' privacy? How was confidentiality maintained? Were Privacy Rule procedures followed (if applicable)? Was a Certificate of Confidentiality obtained? If not, *should* one have been obtained?
8. Were vulnerable groups involved in the research? If yes, were special precautions used because of their vulnerable status?
9. Were groups omitted from the inquiry without a justifiable rationale, such as women (or men), minorities, or older people?

You can also come to some conclusions based on a description of the study methods. There may be sufficient information to judge, for example, whether study participants were subjected to physical or psychological harm or discomfort. Reports do not always specifically state whether informed consent was secured, but you should be alert to situations in which the data could not have been gathered as described if participation were purely voluntary (e.g., if data were gathered unobtrusively).

In thinking about ethical issues, you should also consider who the study participants were. For example, if a study involved vulnerable groups, there should be more information about protective procedures. You might also need to attend to who the study participants were *not*. For example, there has been considerable concern about the omission of certain groups (e.g., minorities) from clinical research.

It is often difficult to determine whether the participants' privacy was safeguarded unless the researcher mentions pledges of confidentiality or anonymity. A situation requiring special scrutiny arises when data are collected from two people simultaneously (e.g., a husband and wife who are jointly interviewed); in such situations, the absence of privacy raises not only ethical concerns, but also questions regarding participants' candor. As noted by Forbat and Henderson (2003), ethical issues arise when two people in an intimate relationship are interviewed about a common issue, even when they are interviewed privately. They described the potential for being "stuck in the middle" when trying to get two sides of a story, and facing the dilemma of how to ask one person probing questions after having been given confidential information about the topic by the other.

RESEARCH EXAMPLES

Two research examples that highlight ethical issues are presented in the following sections.

Research Example from a Quantitative Study

Study: Health status in an invisible population: Carnival and migrant worker children (Kilanowski & Ryan-Wenger, 2007).

Study Purpose: The purpose of the study was to examine the health status of children of itinerant carnival workers and migrant farm workers in the United States.

Research Methods: A total of 97 boys and girls younger than 13 years were recruited into the study. All children received an oral health screening and were measured for height and weight. Parents completed questionnaires about their children's health and healthcare, and most brought health records from which information about immunizations was obtained.

Ethics-Related Procedures: The families were recruited through the cooperation of gatekeepers at farms and carnival communities in 7 states. Parents were asked to complete informed consent forms, which were available in both English and Spanish. Children who were older than 9 were also asked whether they would like to participate, and gave verbal assent. Confidentiality was a concern to both the families and the gatekeepers. The researchers needed to assure all parties that the data would be confidential and not used against families or facilities. Data were gathered in locations and time periods that had been suggested by the carnival managers and farm owners so that parents did not need to forfeit work hours to participate in the study. Migrant farm workers were often eager to participate, and often waited in line to sign the consent forms. At the conclusion of the encounter, the researchers gave the parents a written report of the children's growth parameters and recommendations for follow-up. In appreciation of the parents' time, $10 was given to the parents, and the child was given an age-appropriate nonviolent toy (worth about $10) of their choice. Children were also given a new toothbrush. The IRB of the Ohio State University approved this study.

Key Findings: Carnival children were less likely than migrant children to have regularly scheduled well-child examinations and to have seen a dentist in the previous year. Among children ages 6 to 11, the itinerant children in both groups were substantially more likely to be overweight than same-aged children nationally.

Research Example from a Qualitative Study

Study: Storying childhood sexual abuse (Draucker & Martsolf, 2008).

Study Purpose: The purpose of the study was to describe and explain how individuals disclose their experience of childhood sexual abuse.

Study Methods: Drauker and Martsolf used grounded theory methods to develop a framework explaining how survivors of childhood sexual abuse tell others about their abuse experiences. The study data were from open-ended interviews with 74 individuals (40 women and 34 men) who had experienced ongoing sexual abuse by a family member or close acquaintance. The interviews were audiotaped for subsequent analysis.

Ethics-Related Procedures: Prospective participants were screened before enrollment in the study to ensure that they were not experiencing psychiatric distress or current abuse that would make participation risky. Informed consent was obtained from individuals who passed the screening. Participants were paid $35 for their time and travel expenses. Emergency mental health referral procedures were developed in case a participant experienced acute distress during the interview. No one required an emergency referral, but several people requested information about counseling resources. The researchers obtained IRB approval from their university prior to data collection. A Certificate of Confidentiality was obtained to ensure participants' privacy.

Key Findings: The psychological problem faced by participants was that childhood sexual abuse both demands and defies explanation. The core psychological process used in response to this problem was called "storying childhood sexual abuse." Processes included: (1) starting the story: the story-not-yet-told; (2) coming out with the story: the story-first-told; (3) shielding the story: the story-as-secret; (4) revising the story: the story-as-account; and (5) sharing the story: the story-as-message.

SUMMARY POINTS

- Because research has not always been conducted ethically and because researchers face **ethical dilemmas** in designing studies that are both ethical and rigorous, **codes of ethics** have been developed to guide researchers.
- Three major ethical principles from the *Belmont Report* are incorporated into most guidelines: beneficence, respect for human dignity, and justice.
- **Beneficence** involves the performance of some good and the protection of participants from

physical and psychological harm and exploitation (*nonmaleficence*).

- **Respect for human dignity** involves participants' **right to self-determination,** which means they have the freedom to control their own actions, including voluntary participation.
- **Full disclosure** means that researchers have fully divulged participants' rights and the risks and benefits of the study. When full disclosure could yield biased results, researchers sometimes use **covert data collection** or **concealment** (the collection of information without the participants' knowledge or consent) or **deception** (either withholding information from participants or providing false information).
- **Justice** includes the **right to fair treatment** and the **right to privacy.** In the United States, privacy has become a major issue because of the Privacy Rule regulations that resulted from the Health Insurance Portability and Accountability Act (HIPAA).
- Various procedures have been developed to safeguard study participants rights, including risk/benefit assessments, informed consent procedures, and confidentiality procedures.
- In a **risk/benefit assessment,** the potential benefits of the study to participants and to society are weighed against the costs to individuals.
- **Informed consent** procedures, which provide prospective participants with information needed to make a reasoned decision about participation, normally involve signing a **consent form** to document voluntary and informed participation.
- In qualitative studies, consent may need to be continually renegotiated with participants as the study evolves, through **process consent** procedures.
- Privacy can be maintained through **anonymity** (wherein not even researchers know participants' identities) or through formal **confidentiality procedures** that safeguard the information participants provide.
- U.S. researchers can seek a **Certificate of Confidentiality** that protects them against the forced disclosure of confidential information through a court order or other legal or administrative process.
- Researchers sometimes offer **debriefing** sessions after data collection to provide participants with more information or an opportunity to air complaints.
- **Vulnerable groups** require additional protection. These people may be vulnerable because they are unable to make a truly informed decision about study participation (e.g., children), because of diminished autonomy (e.g., prisoners), or because circumstances heighten the risk of physical or psychological harm (e.g., pregnant women).
- External review of the ethical aspects of a study by an ethics committee, Research Ethics Board (REB), or **Institutional Review Board (IRB)** is highly desirable and may be required by either the agency funding the research or the organization from which participants are recruited.
- In studies in which risks to participants are minimal, an **expedited review** (review by a single member of the IRB) may be substituted for a full board review; in cases in which there are no anticipated risks, the research may be exempted from review.
- Researchers need to give careful thought to ethical requirements throughout the study's planning and implementation and to ask themselves continually whether safeguards for protecting humans are sufficient.
- Ethical conduct in research involves not only protection of the rights of human and animal subjects, but also efforts to maintain high standards of integrity and avoid such forms of **research misconduct** as *plagiarism*, *fabrication* of results, or *falsification* of data.

STUDY ACTIVITIES

Chapter 7 of the *Resource Manual for Nursing Research: Generating and Assessing Evidence for Nursing Practice, 9th ed.*, offers study suggestions for reinforcing concepts presented in this chapter. In addition, the following questions can be addressed in classroom or online discussions:

1. For one of the two studies described in the research example section (Kilanowski and Ryan-Wegner, 2007, or Draucker and Martsolf, 2008), draft a consent form that includes required information, as described in the section on informed consent.
2. Answer the relevant questions in Box 7.3 regarding the Kilanowski and Ryan-Wenger (2007) study. Also consider the following questions: (a) Could the data for this study have been collected anonymously? Why or why not? (b) Might a Certificate of Confidentiality have been helpful in this study?
3. Answer the relevant questions in Box 7.3 regarding the Draucker and Martsolf (2008) study. Also consider the following questions: (a) The researchers paid participants a $35 stipend—was this ethically appropriate? (b) Why do you think the researchers obtained a Certificate of Confidentiality for this research?

STUDIES CITED IN CHAPTER 7

Anderson, G. (2008). Ethical preparedness and performance of gene therapy study coordinators. *Nursing Ethics*, *15*, 208–221.

Artinian, N., Flack, J., Nordstrom, C., Hockman, E., Washington, O., Jen, K., & Fathy, M. (2007). Effects of nurse-managed telemonitoring on blood pressure at 12-month follow-up among urban African Americans. *Nursing Research*, *56*, 312–322.

Caelli, K. (2001). Engaging with phenomenology: Is it more of a challenge than it needs to be? *Qualitative Health Research*, *11*, 273–281.

Caldwell, B., & Redeker, N. (2009). Sleep patterns and psychological distress in women living in the inner city. *Research in Nursing & Health*, *32*, 177–190.

Carlsson, E., Paterson, B., Scott-Findley, S., Ehnfors, M., & Ehrenberg, A. (2007). Methodological issues in interviews involving people with communication impairments after acquired brain damage. *Qualitative Health Research*, *17*, 1361–1371.

Draucker, C. B., & Martsolf, D. (2008). Storying childhood sexual abuse. *Qualitative Health Research*, *18*, 1034–1048.

Freda, M. C., & Kearney, M. (2005). Ethical issues faced by nursing editors. *Western Journal of Nursing Research*, *27*(4), 487–499.

Graffigna, G., & Olson, K. (2009). The ineffable disease: Exploring young people's discourses about HIV/AIDS in Alberta, Canada. *Qualitative Health Research*, *19*, 790–801.

Habermann, B., Broome, M., Pryor, E., & Ziner, K. W. (2010). Research coordinators' experiences with scientific misconduct and research integrity. *Nursing Research*, *59*, 51–57.

Horgas, A., Yoon, S., Nichols, A., & Marsiske, M (2008). The relationship between pain and functional disability in black and white older adults. *Research in Nursing & Health*, *31*, 341–354.

Hsiao, C. Y., & Van Riper, M. (2009). Individual and family adaptation in Taiwanese families of individuals with severe and persistent mental illness. *Research in Nursing & Health*, *32*, 307–320.

Jones, B., Volker, D., Vinajeras, Y., Butros, L., Fitchpatrick, C., & Rossetto, K. (2010). The meaning of surviving cancer for Latino adolescents and emerging young adults. *Cancer Nursing*, *33*, 74–81.

Kelly, P., Cheng, A., Peralez-Dieckmann, E., & Martinez, E. (2009). Dating violence and girls in the juvenile justice system. *Journal of Interpersonal Violence*, *24*, 1536–1551.

Kilanowski, J., & Ryan-Wenger, N. (2007). Health status in an invisible population: Carnival and migrant worker children. *Western Journal of Nursing Research*, *29*, 100–120.

Laughon, K. (2007). Abused African American women's processes of staying healthy. *Western Journal of Nursing Research*, *29*, 365–384.

Polit, D. F., & Beck, C. T. (2009). International gender bias in nursing research, 2005–2006: A quantitative content analysis. *International Journal of Nursing Studies*, *46*, 1102–1110.

Raines, T., Jones, P., Moe, N., Duncan, R., McCall, S., & Caremuga, T. (2009). Investigation of the anxiolytic effects of luteolin, a lemon balm flavonoid in the male Sprague-Dawley rat. *AANA Journal*, *77*, 33–36.

Rothstein, W., & Phuong, L. (2007). Ethical attitudes of nurse, physician, and unaffiliated members of Institutional Review Boards. *Journal of Nursing Scholarship*, *39*, 75–81.

Sandgren, A., Thulesius, H., Fridlund, B., & Petersson, K. (2006). Striving for emotional survival in palliative cancer nursing. *Qualitative Health Research*, *16*, 79–96.

Speraw, S. (2009). "Talk to me—I'm human": The story of a girl, her personhood, and the failures of health care. *Qualitative Health Research*, *19*, 732–743.

Treacy, M., Hyde, A., Boland, J., Whitaker, T., Abaunza, P. S., & Stewart-Knox, B. (2007). Children talking: Emerging perspectives and experiences of cigarette smoking. *Qualitative Health Research*, *17*, 238–249.

Wagner, L., Capezuti, E., & Rice, J. (2009). Nurses' perceptions of safety culture in long-term care settings. *Journal of Nursing Scholarship*, *41*, 184–192.

Methodologic and nonresearch references cited in this chapter can be found in a separate section at the end of the book.

29 Writing Proposals to Generate Evidence

Research proposals communicate a research problem and proposed methods of solving it to an interested party. Research proposals are written both by students seeking faculty approval for studies and by researchers seeking financial support. In this chapter, we offer tips on how to improve the quality of research proposals and how to develop proficiency in **grantsmanship**—the set of skills involved in securing research funding.

OVERVIEW OF RESEARCH PROPOSALS

In this section, we provide some general information regarding research proposals. Most of the information applies equally to dissertation proposals and grant applications.

Functions of a Proposal

Proposals are a means of opening communication between researchers and other parties. Those parties typically are either funding agencies or faculty advisers, whose job it is to accept or reject the proposed plan or to request modifications. An accepted proposal is a two-way contract: those accepting the proposal are effectively saying, "We are willing to offer our (emotional or financial) support, for a study that proceeds as proposed," and those writing the proposal are saying, "If you offer support, then we will conduct the study as proposed."

Proposals often serve as the basis for negotiating with other parties as well. For example, a proposal may be shared with administrators when seeking institutional approval to conduct a study (e.g., for gaining access to participants). Proposals are often incorporated into submissions to human subjects committees or Institutional Review Boards.

Proposals help researchers to clarify their own thinking. By committing ideas to writing, ambiguities can be addressed at an early stage. Proposal reviewers also offer suggestions for conceptual and methodologic improvements. When studies are undertaken collaboratively, proposals can help ensure that all researchers are "on the same page" about how the study is to proceed and can thus minimize the possibility of friction.

Proposal Content

Proposal reviewers want a clear idea of what the researcher plans to study, why the study is needed, what methods will be used to achieve study goals, how and when tasks are to be accomplished, and whether the researcher has the skills to complete the project successfully. Proposals are evaluated on a number of criteria, including the importance of

the question, the adequacy of the methods, and, if money is being requested, the reasonableness of the budget.

Proposal authors are usually given instructions about how to structure proposals. Funding agencies often supply an application kit that includes forms to be completed and specifies the format for organizing proposal contents. Universities issue guidelines for dissertation proposals.

The content and organization of most proposals are broadly similar to that for a research report, but proposals are written in the future tense (i.e., indicating what the researcher *will* do) and obviously do not include results and conclusions. The description of proposed methods—what the researchers propose to do to develop evidence that is valid and trustworthy—is critically important to the success of the proposal.

Proposals for Qualitative Studies

Preparing proposals for qualitative research entails special challenges. Methodologic decisions typically evolve in the field; therefore, it is seldom possible to provide detailed or in-depth information about such matters as sample size or data collection strategies. Sufficient detail needs to be provided, however, so that reviewers will have confidence that the researcher will assemble strong data and do justice to the data collected.

Qualitative researchers must persuade reviewers that the topic is important and worth studying, that they are sufficiently knowledgeable about the challenges of field work and adequately skillful in eliciting rich data, and, in short, that the project would be a very good risk. Knafl and Deatrick (2005) offered 10 tips for successful qualitative proposals. The first tip was to make the case for the *idea*, not the method. Qualitative researchers were also advised to avoid methodologic tutorials, to use examples to clarify the research design, and to write for both the experts and the skeptics.

Resources are available to help qualitative researchers with proposal development. For example, an entire issue of the journal *Qualitative Health Research* was devoted to proposal writing—the July 2003 issue (volume 13, issue 6). Useful advice is also available in Morse and Richards (2002), Sandelowski and colleagues (1989), and Padgett and Henwood (2009).

Proposals for Theses and Dissertations

Dissertation proposals are sometimes a bigger hurdle than dissertations themselves. Many doctoral candidates founder at the proposal development stage rather than when writing or defending the dissertation. Much of our advice—especially in our "Tips" section later in the chapter—applies equally to proposals for theses and dissertations as for grant applications, but some additional advice might prove helpful.

The Dissertation Committee

Choosing the right adviser (if an adviser is chosen rather than appointed) is almost as important as choosing the right research topic. The ideal adviser is one who is a mentor, an expert with a strong reputation in the field, a good teacher, a patient and supportive coach and critic, and an advocate. The ideal adviser is also a person who has sufficient time and interest to devote to your research and who is likely to stick with your project until its completion. This means that it might matter whether the prospective adviser has plans for a sabbatical leave, or is nearing retirement.

Dissertation committees often involve three or more members. If the adviser lacks certain "ideal" characteristics, those characteristics can be balanced across committee members by seeking people with complementary talents. Putting together a group who will work well together and who have no personal antagonism toward each other can, however, be tricky. Advisers can usually make good suggestions about other committee members.

Once a committee has been formed, it is important to develop a good working relationship with members and to learn about their viewpoints before and during the proposal development stage. This means, at a minimum, becoming familiar with their research and the methodologic strategies they have favored. It also means meeting with them and sounding

them out with ideas about topics and methods. If the suggestions from two or more members are at odds, it is prudent to seek your adviser's counsel on how to resolve this.

TIP: When meeting with your adviser and committee members, take notes about their suggestions, and write them out in more detail after the meeting while they are still fresh in your mind. The notes should be reviewed while developing the proposal.

Practices vary from one institution to another and from adviser to adviser, but some faculty require a prospectus before giving the go-ahead to prepare a full proposal. The prospectus is usually a three- to four-page paper outlining the research questions and proposed methods.

Content of Dissertation Proposals

Specific requirements regarding the length and format of dissertation proposals vary in different settings, and it is important to know at the outset what is expected. Typically, dissertation proposals are 20 to 40 pages in length. In some cases, however, committees prefer "mini-dissertations," that is, a document with fully developed sections that can be inserted with minor adaptation into the dissertation itself. For example, the review of the literature, theoretical framework, hypothesis formulation, and the bibliography may be sufficiently refined at the proposal stage that they can be incorporated into the final product.

Literature reviews are often the most important section of a dissertation proposal (at least for quantitative studies). Committees may not desire lengthy literature reviews, but they want to be assured that students are in command of knowledge in their field of inquiry.

Dissertation proposals sometimes include elements not normally found in proposals to funding agencies. One such element may be table shells (see Chapter 19), which can demonstrate that the student knows how to analyze data and present results effectively. Another element is a table of contents for the dissertation. The table of contents serves as an outline for the final product, and shows that the student knows how to organize material.

Several books provide additional advice on writing a dissertation proposal, including Locke and colleagues (2007) and Rudestam and Newton (2007).

FUNDING FOR RESEARCH PROPOSALS

Funding for research projects is becoming increasingly difficult to obtain because of keen and growing competition. As more nurses gain research skills, and as the push for evidence-based practice grows, so too are applications for research funding increasing. Successful proposal writers need to have good research and proposal-writing skills, and they must also know from whom funding is available.

Federal Funding in the United States

The largest funder of research activities in the United States is the federal government. For healthcare researchers, the National Institutes of Health (NIH) and the Agency for Healthcare Research and Quality (AHRQ) are leading agencies. Two major types of federal disbursements are grants and contracts. **Grants** are awarded for studies conceived by researchers themselves, whereas **contracts** are for studies desired by the government.

There are several mechanisms for NIH grants, which can be awarded to researchers in both domestic and foreign institutions. Most grant applications are unsolicited, and reflect the research interests of individual researchers. Unsolicited applications should be consistent with the broad objectives of an NIH funding agency, such as NINR. Investigator-initiated applications are submitted in response to **Parent Announcements**, which are covered under omnibus Funding Opportunity Announcements (FOAs).

NIH also issues periodic **Program Announcements** (PA) that describe new, continuing, or expanded program interests. For example, in March 2010, NINR issued a joint program announcement with 16 other NIH institutes titled "Behavioral and Social Science Research on Understanding and Reducing Health Disparities" (PA-10-136). The purpose of this PA, which expires in 2013, is "to encourage behavioral and social science research

on the causes and solutions to health and disabilities disparities in the U.S. population."

Another grant mechanism allows federal agencies to identify a *specific* topic area in which they are interested in receiving proposals by a **Request for Applications** (**RFA**). RFAs are one-time opportunities with a single submission date. As an example, NINR issued an RFA titled "Chronic Co-Morbid Conditions in HIV+ U.S. Adults on Highly-Effective Anti-Retroviral Therapy" in February 2010, with grant applications due in May 2010. The RFA states general guidelines and goals for the competition, but researchers can develop the specific research problem within the topic area. A weekly electronic publication, the *NIH Guide for Grants and Contracts*, contains announcements about RFAs, PAs, and Parent Announcements.

In addition to grants, some government agencies award contracts to do *specific* studies. Contract offers are announced in a **Request for Proposals** (**RFP**), which details the *exact* study that the government wants. Contracts, which are typically awarded to only one competitor, constrain researchers' activities and so most nurse researchers compete for grants rather than contracts. A summary of federal RFPs is published in the *Commerce Business Daily* (*http://cbdnet.gpo.gov*).

Government funding for nursing research is, of course, also available in other countries. In Canada, for example, various types of health research are sponsored by the Canadian Institutes of Health Research (CIHR). Information about CIHR's program of grants, training awards, and other funding opportunities is available at its website (*http://www.cihr.ca*).

Private Funds

Healthcare research is supported by numerous philanthropic foundations, professional organizations, and corporations. Many researchers prefer private funding to government support because there is less "red tape" and fewer requirements.

Information about philanthropic foundations that support research is available through the Foundation Center (*http://www.fdncenter.org*). A comprehensive resource for identifying funding opportunities is the Center's *The Foundation Directory*, now available online for a fee. The directory lists the purposes and activities of the foundations and information for contacting them. The Foundation Center also offers seminars and training on grant writing and funding opportunities in locations around the United States. Another resource for information on funding is the Community of Science, which maintains a database on funding opportunities (*http://www.cos.com*).

Professional associations (e.g., the American Nurses' Foundation, Sigma Theta Tau) offer funds for conducting research. Health organizations, such as the American Heart Association and the American Cancer Society, also support research activities.

Finally, research funding is sometimes donated by private corporations, particularly those dealing with healthcare products. The Foundation Center publishes a directory of corporate grantmakers and provides links through its website to a number of corporate philanthropic programs. Additional information concerning corporate requirements and interests should be obtained either from the organization directly or from staff in the research administration offices of the institution with which you are affiliated.

GRANT APPLICATIONS TO NIH

NIH funds many nursing studies through NINR and through other institutes. Because of the importance of NINR as a funding source for nurse researchers, this section describes the process of proposal submission and review at NIH. AHRQ, which also funds nurse-initiated studies, uses the same application kit and similar procedures.

Types of NIH Grants and Awards

NIH awards different types of research grants, and each has its own objectives and review criteria. The basic grant program—and the primary funding mechanism for independent research—is the traditional **Research Project Grant** (R01). The objective of R01 grants is to support specific research projects in areas reflecting the interests and competencies of a Principal Investigator (PI).

Beside the R01 grant program, three others that are available through NINR are worth noting. A special program (R15) has been established for researchers working in institutions that have not been major participants in NIH programs. These **Academic Research Enhancement Awards (AREA)** are designed to stimulate research in institutions that provide baccalaureate training for many individuals who go on to do health-related research. There is also a **Small Grant** Program (R03) that provides support for pilot, feasibility, and methodology development studies. R03 grants provide a maximum of $50,000 of direct support for up to 2 years. Finally, the R21 grant mechanism—the **Exploratory/Developmental Research Grant Award**—is intended to encourage new, exploratory, and developmental research projects by providing support for early stages of research.

NIH and other agencies also offer individual and institutional predoctoral and postdoctoral fellowships, as well as career development awards. Examples of individual fellowship mechanisms available through the National Research Service Award (NRSA) program within NINR include the following:

- F31, Ruth Kirshstein Individual Predoctoral NRSA Fellowships, support nurses in a supervised training leading to a doctoral degree in areas related to the NINR mission
- F32, Ruth Kirshstein Individual Postdoctoral NRSA Fellowships, support postdoctoral training to nurses to broaden their scientific background
- F33, Senior NRSA Fellowships, support doctorally trained researchers with at least 7 years of research in pursuing opportunities to change the direction of their research careers.

TIP: Advice on developing a proposal for an NRSA fellowship has been offered in a paper by Parker and Steeves (2005).

Four important Career Development Awards offered through NINR are as follows:

- K01, Mentored Research Scientist Development Award, available to doctorally prepared scientists who would benefit from a mentored research experience with an expert sponsor
- K22, NINR's Career Transition Awards, offers support to postdoctoral fellows in transition to a faculty position
- K23, Mentored Patient-Oriented Research Career Development Award, supports the career development of investigators who are committed to focusing on patient-oriented research
- K99, Pathways to Independence Awards, provide for postdoctoral research activity leading to the submission of an independent research project application.

TIP: If you have an idea for a study and are not sure which type of grant program is suitable—or you are unsure whether NINR or another NIH institute might be interested—you should contact NINR directly (telephone number: 301-594-6906). NINR staff can provide feedback about whether your proposed study matches NINR's program interests. Information about NINR's ongoing priorities and areas of opportunity is available at *http://www.nih.gov/ninr*. A one- to two-page concept paper can also be e-mailed to the address listed on the NINR website.

NIH Forms and Schedule

In 2007, NIH transitioned from hard-copy application submissions to electronic submissions using the SF424 (R&R) application, most recently revised in early 2010, through *www.grants.gov.* The SF424 is used for all the types of grants and awards described in the previous section, although there are supplemental components needed for some of them. Researchers use Adobe Reader (version 8.1.6 or later) to "fill in" and complete this new application. There is abundant information online about the new application process, and NIH offers training sessions on how to submit applications electronically. The application kit can be accessed from the NIH website at *http:// www.nih.gov* under their "Grants and Opportunities" section.

New grant applications are usually processed in three cycles annually. Different deadlines apply to different types of grants, as shown in Table 29.1. For most new applications, except fellowships in the F series and AIDS-related research, the deadline for

TABLE 29.1 Schedule for Selected New Research Applications, National Institutes of Health*

Application Deadline	MECHANISM OF SUPPORT (TYPE OF AWARD)				
	R01	R03, R21	R15	K Series	F Series
Cycle I[a]	February 5	February 16	February 25	February 12	April 8
Cycle II[b]	June 5	June 16	June 25	June 12	August 8
Cycle III[c]	October 5	October 16	October 25	October 12	December 8

[a]Cycle I: Scientific Review: June–July; Earliest start date: December
[b]Cycle II: Scientific Review: October–November; Earliest start date: April
[c]Cycle III: Scientific Review: February–March; Earliest start date: July
*AIDS-related applications are on a different schedule; consult the NIH website for information.

receipt is in February, June, and October. The scientific merit review dates are about 4 to 5 months after each submission date. For example, applications submitted for the February cycle are reviewed in June or July; the earliest project start date for applications funded in that cycle would be in December. Applicants should begin a registration process through the Electronic Research Administration (eRA) Commons at least 2 weeks prior to the submission date.

Preparing a Grant Application for NIH

Although many substantive aspects of the NIH grant application have remained stable, the forms and procedures for NIH grant applications have been changing. It is crucial to carefully review up-to-date instructions for grant application submission rather than relying on information in this chapter.

Forms: Screens and Uploaded Attachments

The SF424 form set has numerous components. The "front matter" of SF424 consists of various forms that appear on a series of fillable screens. These forms help in processing the application and provide administrative information. Careful attention to detail with these forms is very important. Major forms include the following:

- *Cover Component.* On the cover form, researchers state a brief, descriptive title of the project (not to exceed 81 characters), the name and affiliation of the PI, and other administrative information.

TIP: The project title should be given careful thought. It is the first thing that reviewers see, and should be crafted to create a good impression. The title should be concise and informative, but should also be compelling.

- *Project/Performance Site Location Component.* The next screen requests information about the primary site where the work will be performed.
- *Other Project Information Component.* This screen is the mechanism for submitting key information. The form begins with questions about human subjects, and the last few items require attachments to be uploaded, including a project summary, a project narrative, bibliography, and facilities and equipment information. Attachments, which must be in PDF format, have strict size limitations. The *Project Summary* serves as a succinct description of aims and methods of the proposed study and must be no longer than 30 lines. The *Project Narrative* is a brief (two to three sentences) description of the relevance of the research to public health. The *Bibliography* is a list of references cited in the research plan; any reference style is acceptable.

The *Facilities* attachment is used to describe needed and available resources (e.g., laboratories).

- *Senior/Key Person Profile(s) Component.* For each key person, the form requests basic identifying information and calls for an attachment, a Biographical Sketch. The sketch must list education and training, as well as the following: (a) a personal statement describing the qualifications that make the person well suited for his or her role, (b) positions and honors, (c) selected peer-reviewed publications or manuscripts in press, and (d) recently completed and ongoing research support. A maximum of four pages is permitted for each person.
- *Budget Component*. For NIH applications, researchers must chose between two budget options—the R&R Budget Component or the PHS398 Modular Budget Component. Detailed R&R budgets showing specific projected expenses are required if annual direct project costs exceed $250,000, but for smaller projects, budget information is obtained in another section. (Modular budgets are only appropriate for R-type grants.)

For grant applications to NIH and other public health service agencies, additional forms referred to as PHS398 components are required and include the following:

- *Cover Letter Component*. Cover letters to the funding agency are strongly encouraged. Information in the cover letter should include the application title, the name and number of the funding opportunity, and any request to be assigned to a particular review group.
- *Cover Page Supplement Component.* This form supplements the SF424 cover page and requests mainly administrative information.
- *Modular Budget Component.* **Modular budgets**, paid in modules of $25,000, are appropriate for R-series applications (e.g., R01s) requesting $250,000 or less per year of direct costs. (**Direct costs** include specific project-related costs such as staff and supplies; **indirect costs** are institutional **overhead** costs.) This form provides budget fields for annual summaries of projected costs for up to 5 years of support. There are also fields for cumulative summaries across all project years. A *budget justification* attachment, detailing primarily personnel costs, must be uploaded.

TIP: Even though modular budget forms ask only for summaries of the funds needed to complete a study, you should prepare a more detailed budget to arrive at a reasonable projection of needed funds. Beginning researchers are likely to need the assistance of a research administrator or an experienced, funded researcher in preparing their first budget. Higdon and Topp (2004) have offered some advice on developing a budget.

- *Research Plan Component.* The PHS398 Research Plan form asks about application type (e.g., new, resubmission) and then requires information, in the form of attachments, about the proposed study and the research plan. Research plan requirements are described in the next section.
- *Checklist Component*. The checklist includes various miscellaneous items, including organizational assurances and certifications.

TIP: Examples of selected forms for SF424 are presented in the Toolkit of the *Resource Manual* in nonfillable form—that is, they are included simply as illustrations, not to be used for submitting a grant application.

The Research Plan Component

The research plan component consists of 16 items, not all of which are relevant to every application—for example, item 1 is for revised applications or resubmissions. Each item involves uploading separate PDF attachments. In this section, we briefly describe guidelines for items 2 through 16, with emphasis on items 2 and 3. We also present some advice based on a study (Inouye & Fiellin, 2005) in which the researchers content-analyzed the criticisms in the review sheets of 66 R01 applications submitted to a clinical research review group (not NINR). Thus, the advice relating to specific pitfalls is "evidence-based," that is, based on identified problems in actual applications.

TIP: Based on their analysis, Inouye and Fiellin (2005) created a grant-writing checklist designed as a self-assessment tool for proposal developers. We have included an adapted and expanded checklist in the Toolkit that is part of the accompanying *Resource Manual.*

Specific Aims (Item 2). In this section, which is restricted to a single page, researchers must provide a succinct summary of the research problem and the specific objectives of the study, including any hypotheses to be tested. The aims statement should indicate the scope and importance of the problem. Care should be taken to be precise and to identify a problem of manageable proportions—a broad and complex problem is unlikely to be solvable.

Inouye and Fiellin (2005) found that the most frequent critique of the Specific Aims section was that the goals were overstated, overly ambitious, or unrealistic (18% of the review sheets). Other complaints were that the project was poorly conceptualized (15%) or that hypotheses were not clearly articulated (12%).

Research Strategy (Item 3). In the new application forms released in 2010, several sections from earlier forms (Background, Preliminary Studies, and Research Design and Methods) were combined and page restrictions were severely tightened. Unless otherwise specified in a Funding Opportunity Announcement (FOA), the Research Strategy section is now restricted to 12 pages for R01 and R15 applications, and to 6 pages for R03, R21, and F-series applications. For other funding mechanisms, page restrictions are specified in the FOA.

TIP: Career Development Awards (K-series) involve completion of a special form, requiring attachments that include a description of the applicant's background, a statement of career goals and objectives, career development or training activities during the award period, and training in the responsible conduct of research. These items plus the Research Strategy section must, in combination, be no more than 12 pages. The applicant's institution must also submit a letter describing its commitment to the candidate and to his or her development.

The Research Strategy section is organized into three subsections: Significance, Innovation, and Approach. In the Significance section, researchers must convince reviewers that the proposed study idea has clinical or theoretical relevance and that the study will make a contribution to scientific knowledge or clinical practice. Researchers describe the study context in this section through a brief analysis of existing knowledge and gaps on the topic. Researchers should demonstrate command of current knowledge in a field, but this section must be very tightly written. Inouye and Fiellin (2005) found that a frequent critique expressed by reviewers about this section was that the need for the study was not adequately justified (29%). In the Innovation section, researchers should describe how the proposed study challenges, refines, or improves current research or clinical practice paradigms.

The proposed design and methods for the study are described in the third subsection, Approach. This section, which is the heart of the application, should be written with extreme care and reviewed with a self-critical eye. The Approach section needs to be concise, but with sufficient detail to persuade reviewers that methodologic decisions are sound and that the study will yield important and reliable evidence.

A thorough Approach section typically describes the following: (1) the research design, including a discussion of comparison group strategies and methods of controlling confounding variables (for qualitative studies, the research tradition should be described); (2) the experimental intervention, if applicable, including a description of the treatment and control group conditions; (3) procedures, such as what equipment will be used, how participants will be assigned to groups, and what type of blinding, if any, will be achieved; (4) the sampling plan, including eligibility criteria and sample size; (5) data collection methods and information about reliability and validity of measures; and (6) data analysis strategies. The Approach should identify potential methodologic problems and intended strategies for handling such problems. In proposals for qualitative studies, special care should be given to steps that will be taken to enhance the integrity and trustworthiness of the study.

Inouye and Fiellin (2005) found that *all* of the reviews they analyzed had one or more criticism of this section, the most general of which was that the description of methods was underdeveloped (15%). A few of the most persistent criticisms were as follows:

- Inadequate blinding for outcome assessment (36%)
- Sample was flawed—biased or unrepresentative (36%)
- Important confounding variables inadequately controlled (32%)
- Inadequate sample size or inadequate power calculations (26%)
- Insufficient description of the approach to data analysis (24%)
- Outcome measures inadequately specified or described (23%)

Although some of these concerns relate to clinical trials (e.g., blinding), many have broad relevance—small sample size, sample biases, uncontrolled variables, and poorly described data collection and analysis plans can be problematic in any type of study.

The Approach section must also include information on Preliminary Studies. In new applications, researchers must describe the PI's preliminary or developmental studies and any experience pertinent to the application. This section must persuade reviewers that you have the skills and background needed to do the research. Any pilot work that has served as a foundation for the proposed project should be described. Inouye and Fiellin's (2005) analysis is especially illuminating with regard to Preliminary Studies. They found that the single biggest criticism across the 66 review sheets was that more pilot work was needed, mentioned in 41% of the reviews.

TIP: For applications submitted by Early Stage Investigators (a person within 10 years of completing their terminal degree and who has not yet been awarded an R01 grant), reviewers are instructed to place less emphasis on the applicant's Preliminary Studies.

Human Subjects Sections (Items 6–9). Researchers who plan to collect data from human subjects must complete items relating to the protection of subjects. An entire section of the application kit ("Part II, Supplemental Instructions for Preparing the Human Subjects Section of the Research Plan") provides guidance on the attachments needed for these items. Applicants must either address the involvement of human subjects and describe protections from research risks or provide a justification for exemption with enough information that reviewers can determine the appropriateness of requests for exemption. If no exemption is sought, the section must address various issues, as outlined in the application kit. The application must also include various types of information regarding the inclusion of women, minorities, and children. These sections often serve as the cornerstone of the document submitted to Institutional Review Boards.

Other Research Plan Sections (Items 10–15). Most remaining sections in the research plan component are not relevant universally. These include such items as a description and justification of the use of vertebrate animals and a leadership plan if there are multiple principal investigators. One item, however, has relevance to many applications: Letters of support (Item 14). This item requires you to attach letters from individuals agreeing to provide services to the project, such as consultants.

Appendices (Item 16). Grant applications often include appended materials. A maximum of 10 PDF attachments is allowed, and a summary sheet listing all appended items is encouraged. Examples of appended materials include data collection instruments, clinical protocols, detailed sample size calculations, complex statistical models, and other supplementary materials in support of the application. Researchers can no longer submit publications or manuscripts, except under restricted circumstances. Essential information should never be relegated to an appendix because only primary reviewers receive appendices. The guidelines warn that appendices should not be used to circumvent the page limitations of the Research Strategy section.

TIP: In terms of content, the research plan for NIH applications is similar to what is required in most research proposals—although emphases and page restrictions may vary, and supplementary information may be required.

The Review Process

Grant applications submitted to NIH are reviewed for completeness and relevance by the NIH Center for Scientific Review. Acceptable applications are assigned to an appropriate Institute or Center, and to a peer review group.

NIH uses a sequential, dual review system for informing decisions about its grant applications. The first level involves a panel of peer reviewers (not NIH employees), who evaluate applications for their scientific merit. These review panels are called **scientific review groups** (SRGs) or, more commonly, **study sections**. Each panel consists of about 20 researchers with backgrounds appropriate to the specific study section for which they have been selected. Appointments to the review panels are usually for 4-year terms and are staggered so that about one-fourth of each panel is new each year.

TIP: Applications by nurse researchers usually are assigned to one of two Nursing Science study sections. One is the "Nursing Science: Adults and Older Adults Study Section" (NSAA) and the other is the "Nursing Science: Children and Families Study Section" (NSCF). Fellowship applications in the F series are reviewed in a separate study section, often with K-series applications.

The second level of review is by a National Advisory Council, which includes scientific and lay representatives. The Advisory Council considers not only the scientific merit of an application but also the relevance of the proposed study to the programs and priorities of the Center or Institute to which the application has been submitted, as well as budgetary considerations.

Applications are assigned to primary and secondary (and sometimes tertiary) reviewers for detailed analysis. Each assigned reviewer prepares comments and assigns scores according to five core review criteria.

1. *Significance*. Does the proposed study address an important problem? If the aims of the application are achieved, how will scientific knowledge or clinical practice be advanced? What will be the effect of the study on the concepts or methods that drive this field?
2. *Investigator*. Is the investigator appropriately trained and well suited to carry out this work? Is the proposed work appropriate to the experience level of the PI and other researchers? Do Early Stage Investigators have appropriate training and experience?
3. *Innovation*. Does the project employ novel concepts, approaches, or methods? Are the aims original and innovative? Does the project challenge existing paradigms or develop new methods or technologies?
4. *Approach*. Are the overall strategy, design, methods, and analyses adequately developed and appropriate to the aims of the project? Does the applicant acknowledge potential problem areas and consider alternative tactics?
5. *Environment*. Does the scientific environment in which the work will be done contribute to the probability of success? Do the proposed experiments take advantage of unique features of the scientific environment or employ useful collaborative arrangements? Is there evidence of institutional support?

In addition to these five criteria, other factors are relevant in evaluating proposals, including the reasonableness of the proposed budget, the adequacy of protections for human or animal subjects, and the appropriateness of the sampling plan in terms of including women, minorities, and children as participants. These factors are not, however, formally scored.

Scoring of applications changed in 2010. In the current system, each of the five core criteria is scored on a scale from 1 (exceptional) to 9 (poor). Assigned reviewers score applications and submit their scores before attending a study section

meeting, and also submit a preliminary overall **impact score** (also called a **priority score**) on the same 1 to 9 scale. An impact score reflects a reviewer's assessment of the extent to which the study will exert a powerful influence in an area of research. Based on preliminary impact scores, applications with unfavorable scores (usually those in the lower half) are not discussed or scored by the entire study section in its meeting. This streamlined process was instituted so that study section members could focus their discussion on the most worthy applications.

For applications that *are* discussed in the meeting, each study section member (not just those who were assigned as reviewers) designates an impact score, based on their own critique of the application and the committee's discussion. Individual impact scores from all committee members are averaged, and the mean is then multiplied by 10 to arrive at a final score. Thus, final impact scores for applications that are discussed can range from 10 (the best possible score) to 90 (the lowest possible score). Final scores tend to cluster in the 10 to 50 range, however, inasmuch as the least meritorious applications were previously screened out and not scored by the full study section. Among all scored applications, only those with the best priority scores actually obtain funding. Cut-off scores for funding vary from agency to agency and year to year, but a score of 20 or lower may be needed to secure funding.

Within a few days after the study section meeting, applicants are able to learn their priority score and percentile ranking online via the NIH eRA Commons (*https://commons.era. nih.gov/commons*). Within about 30 days, applicants can access a summary of the study section's evaluation. These **summary sheets** include critiques written by the assigned reviewers, a summary of the study section's discussion, study section recommendations, and administrative notes of special consideration (e.g., human subjects issues). All applicants receive a summary sheet, even if their applications were unscored. (Applicants of unscored applications also learn how the assigned reviewers scored the five core criteria).

TIP: Unless an unfunded proposal is criticized in some fundamental way (e.g., the problem area was not judged to be significant), applications often should be resubmitted, with revisions that reflect the concerns of the peer reviewers. When a proposal is resubmitted, the next review panel members are given a copy of the original application and the summary sheet so that they can evaluate the degree to which initial reviewers' concerns have been addressed. Applications can be resubmitted up to two times.

TIPS ON PROPOSAL DEVELOPMENT

Although it is impossible to tell you exactly what steps to follow to produce a successful proposal, we conclude this chapter with some advice that might help to improve the process and the product. Many of these tips are especially relevant for those preparing proposals for funding. Further suggestions for writing effective grant applications may be found in Beitz and Bliss (2005), Grey (2000), Lusk (2004), and Inouye and Fiellin (2005).

Things to Do before Writing Begins

Advance planning is essential to the development of a successful proposal. This section offers suggestions for things you can do to prepare for the actual writing.

Start Early

Writing a proposal, and attending to all of the details of a formal submission process, is time consuming and almost always takes longer than originally envisioned. Be sure to build in enough time that the product can be reviewed and re-reviewed by members of the team (including any faculty mentors) and by willing colleagues. Make sure there is adequate time for administrative issues such as securing permissions and getting budgets approved.

Having a proposal timeline is a good way to impose discipline on the proposal development process. Figure 29.1 presents one example, but the list of tasks is merely suggestive. Ask an experienced person to review your timeline, and try to adhere to the timeline once you start.

Task	Timeline (Months Before Submission)												
	12+	12	11	10	9	8	7	6	5	4	3	2	1
Identify/conceptualize the problem	X												
Undertake a literature review	X												
Identify and approach possible data collection sites	X												
Initiate descriptive or pilot work	X												
Analyze pilot data, assess feasibility	XXXXX												
Develop a "brief," outlining significance & preliminary thoughts about overall study design		XX											
Identify methodologic and content experts; solicit input and possible collaboration		XXX											
Begin building a team of co-investigators and consultants		XXXX											
Identify contact funder/program officer (as needed)		XX											
Obtain all application forms and instructions		XX											
Review funding agencies' priorities; review recently funded grants		XXX											
Develop research plan, identify instruments, etc.; consult with statisticians, psychometricians, etc., as needed			XXXXXXX										
Collect site data for describing site, staff, clients			XXX										
Obtain written letters of agreement and/or support from data collection sites			XXX										
Prepare an outline of the proposal; develop writing assignments					XX								
Write draft of proposal						XXXXXXX							
Draft a budget							XX						
Draft other ancillary components (bio sketches, etc.)								XX					
Internal review by team members									XXX				
Make revisions based on review										XXX			
External review by colleagues/experts										XXX			
Team review of comments, make final revisions											XXX		
Write abstract/summary												XX	
Finalize budget and other ancillary components												X	
Prepare all final documents, get needed signatures													X

FIGURE 29.1 Example of a grant-writing timeline.

TIP: It is advantageous to build pilot or preliminary work into your proposal development timeline. As noted earlier, NIH reviewers frequently criticize the absence of adequate pilot work. Incremental knowledge building is attractive to reviewers. When you apply for funding, you are asking funders to make an *investment* in you; they will have the sense of being offered a better investment opportunity if some groundwork for a study has already been completed.

Select an Important Problem

A factor that is critical to the success of a proposal is selecting a problem that has clinical or theoretical significance and that is viewed in a positive light by reviewers. The proposal must make a persuasive argument that the research could make a noteworthy contribution to evidence on a topic that is important and appealing to those making recommendations.

Kuzel (2002), who shared some lessons about securing funding for a qualitative study, noted that researchers could profit by taking advantage of certain "hot topics" that have the special attention of the public and government officials. Proposals can sometimes be cast in a way that links them to topics of national concern, and such a linkage can contribute to a favorable review. Kuzel used as an example his funded study of quality of care and medical errors in primary care practices, with emphasis on patient perspectives. The proposal was submitted at a time when the U.S. government was putting resources into research to enhance patient safety and noted

that "the reframing of 'quality' under the name of 'patient safety' has captured the stage and is likely to have an enduring effect on what work receives funding" (p. 141). Both qualitative and quantitative researchers should be sensitive to political realities.

Know Your Audience

Learn as much as possible about the audience for your proposal. For dissertations, this means getting to know your committee members and learning about their expectations, interests, and schedules. If you are writing a proposal for funding, you should obtain information about the funding organization's priorities. It is also wise to examine recently funded projects. Funding agencies often publish the criteria that reviewers use to make funding decisions—such as the ones we described for NIH—and these criteria should be studied carefully.

Grey (2000), in her tips on grantsmanship, urged researchers to "talk it up" (p. 91), that is, to call program staff in agencies and foundations, or to send letters of inquiry about possible interest in a project. Grey also noted the importance of *listening* to what these people say and following their recommendations.

Another aspect to "knowing your audience" concerns appreciating reviewers' perspectives. Reviewers for funding agencies are busy professionals who are taking time away from their own work to consider the merits of proposed new studies. They are likely to be methodologically sophisticated and experts in *their* field—but they may have limited knowledge of your own area of research. It is, therefore, imperative to help time-pressured reviewers to grasp the merits of your proposed study, without relying on jargon or specialized terminology.

Review a Successful Proposal

Although there is no substitute for actually writing a proposal as a learning experience, novice proposal writers can profit by examining a successful proposal. It is likely that some of your colleagues or fellow students have written a proposal that has been accepted (either by a funding sponsor or by a dissertation committee), and many people are glad to share their successful efforts with others. Also, proposals funded by the government are usually in the public domain—that is, you can ask for a copy of funded proposals. To obtain a funded NIH project, for example, you can contact the NIH Freedom of Information Coordinator for the appropriate institute.

Several journals have published entire proposals, except for administrative and budgetary information. An early example was a proposal for a study of comprehensive discharge planning for the elderly (Naylor, 1990). More recently, a proposal for a qualitative study of adolescent fathers was published, together with reviewers' comments (Dallas et al., 2005a, 2005b).

TIP: The accompanying *Resource Manual* includes the entire successful grant application to NINR by Deborah Dillon McDonald entitled "Older adults response to healthcare practitioner pain communication," together with reviewers' comments and McDonald's response.

Create a Strong Research Team

For funded research, it is important to think strategically in putting together a team because reviewers often give considerable weight to researchers' qualifications. It is not enough to have a team of competent people; it is necessary to have the right *mix* of competence. Gaps and weaknesses can often be compensated for by the judicious use of consultants.

Another shortcoming of some project teams is that there are too many researchers with small time commitments. It is unwise to propose a staff with five or more top-level professionals who are able to contribute only 5% to 10% of their time to the project. Such projects often run into management problems because no one is in control of the workflow. Although collaborative work is commendable, you should be able to justify the inclusion of every person.

Things to Do as You Write

If you have planned well and drafted a realistic schedule, the next step is to move forward with the development of the proposal. Some suggestions for the writing stage follow.

Build a Persuasive Case

In a proposal, whether or not funding is sought, you need to persuade reviewers that you are asking the right questions, that you are the right person to ask those questions, and that you will get valid and credible answers. You must also convince them that the answers will make a difference to nursing and its clients.

Beginning proposal writers sometimes forget that they are *selling* a product: themselves and their ideas. It is appropriate, therefore, to think of the proposal as a marketing opportunity. It is not enough to have a good idea and sound methods—you must have a persuasive presentation. When funding is at stake, the challenge is greater because *everyone is trying to persuade reviewers that their proposal is more meritorious than yours.*

Reviewers know that most applications they review will *not* get funded. For example, in fiscal year 2009, fewer than one out of five R01 applications got NIH support. The reviewers' job is to identify the most scientifically worthy applications. In writing the proposal, you must consciously include features that will put your application in a positive light. That is, you should think of ways to gain a competitive edge. Be sure to give thought to issues persistently identified as problematic by reviewers (Inouye & Fiellin, 2005), and use a well-conceived checklist to ensure that you have not missed an opportunity to strengthen your study design and your proposal.

The proposal should be written in a positive, confident tone. If you do not sound convinced that the proposed study is important and will be rigorously done, then reviewers will not be persuaded either. It is unwise to promise what cannot be achieved, but you should think about ways to put the proposed project in a positive light.

Justify Methodologic Decisions

Many proposals fail because they do not instill confidence that key decisions have a good rationale. Methodologic decisions should be made carefully, keeping in mind the benefits and drawbacks of alternatives, and a compelling—if brief—justification should be provided. To the extent possible, make your decisions evidence-based and *defend* the proposed methods with citations demonstrating their utility. Insufficient detail and scanty explanation of methodologic choices can be perilous, although page constraints often make full elaboration impossible.

Begin and End with a Flourish

The abstract or summary to the proposal should be crafted with extreme care. Because it is one of the first things that reviewers read, you need to be sure that it will create a favorable impression. (For NIH applications, nonassigned reviewers may read *only* the summary and not the entire application). The ideal abstract is one that generates excitement and inspires confidence in the proposed study's rigor. Although abstracts appear at the beginning of a proposal, they are often written last.

Proposals typically conclude with material that is somewhat unexciting, such as a data analysis plan. A brief, upbeat concluding paragraph that summarizes the significance and innovativeness of the proposed project can help to remind reviewers of its potential to contribute to nursing practice and nursing science.

Adhere to Instructions

Funding agencies (and universities) provide instructions on what is required in a research proposal. It is crucial to read these instructions carefully and to follow them precisely. Proposals are sometimes rejected without review if they do not adhere to such guidelines as minimum font size or page limitations.

Pay Attention to Presentation

Reviewers are put in a better frame of mind if the proposals they are reading are attractive, well organized, grammatical, and easy to read. Glitzy figures are not needed, but the presentation should be professional and show respect for weary reviewers. In Inouye and Fiellin's (2005) study, 20% of the grant applications were criticized for such presentation issues as typographical or grammatical errors, poor layout, inconsistencies, and omitted tables.

Have the Proposal Critiqued

Before formal submission of a proposal, a draft should be reviewed by others. Reviewers should be selected for both substantive and methodologic

expertise. If the proposal is being submitted for funding, one reviewer ideally would have first-hand knowledge of the funding source. If a consultant has been proposed because of specialized expertise that you believe will strengthen the study, he or she should be asked to participate by reviewing the draft and making recommendations for its improvement.

In universities, mock review panels are often held before submitting a proposal to a funding agency. Faculty and students are invited to these mock reviews and provide valuable feedback for enhancing a proposal.

RESEARCH EXAMPLES

NIH makes available the abstracts of all funded projects through its Research Portfolio Online Reporting Tools (RePORT). Abstracts can be searched by subject, researcher, study section, type of funding mechanism, year of support, and so on. Abstracts for two projects funded through NINR are presented here.

Example of a Funded Quantitative (R01) Project

Elizabeth Schlenk of the University of Pittsburgh prepared the following abstract for a project entitled "Promoting Physical Activity in Older Adults with Comorbidity." The application was reviewed by the Adults and Older Adults Study Section (NSAA), and received NINR funding in March 2010. The project is scheduled for completion in January 2014.

Project Summary: Over 9 million Americans have symptomatic osteoarthritis (OA) of the knee, a chronic disease associated with frequent joint pain, functional limitations, and quadriceps weakness that intrude on everyday life. At least half of those with OA of the knee are diagnosed with hypertension or high blood pressure (HBP), one of the most prevalent risk factors for cardiovascular disease. Many other individuals with OA of the knee unknowingly have HBP and remain untreated. Our own work and that of others suggest that persons with OA of the knee experience reductions in BP when they participate in a regular regimen of physical activity. Even small decreases in systolic and diastolic BP found with physical activity are clinically significant, e.g., a 2 mm Hg decrease reduces the risk of stroke by 14%–17%, and the risk of coronary heart disease is reduced by 6%–9%. Yet, only 15% of persons with OA and 47% with HBP engage in regular physical activity. The purpose of this study is to investigate how the individually delivered, home-based, 6-month modified Staying Active with Arthritis (STAR) intervention, guided by self-efficacy theory and modified to address comorbid HBP, affects lower extremity exercise (flexibility, strengthening, and balance), fitness walking, functional status, BP, quadriceps strength, pain, and health-related quality of life (HRQoL) in a convenience sample of 224 adults age 50 years or older with OA of the knee and HBP. Using a randomized controlled, 2-group design, we (1) hypothesize that at the end of the 6-month intervention period and 6 months after the intervention period ends, those who receive the modified STAR intervention will be more likely to perform lower extremity exercise, participate in fitness walking, show improvements in objective functional status, and demonstrate reductions in BP than those who receive attention-control. Secondarily, we will (2) evaluate the impact of the modified STAR intervention, compared to attention-control, on subjective functional status, quadriceps strength, pain, and HRQoL at both time points; (3) explore the impact of the modified STAR intervention, compared to attention-control, on self-efficacy and outcome expectancy at both time points; (4) explore the relationship between self-efficacy and outcome expectancy; and (5) explore the extent to which self-efficacy and outcome expectancy mediate the relationship between the modified STAR intervention and performance of lower extremity exercise and participation in fitness walking. Data will be analyzed using repeated measures modeling. PUBLIC HEALTH RELEVANCE: The proposed study is relevant to public health because it examines the modified Staying Active with Arthritis (STAR) program to improve leg exercise, fitness walking, and clinical outcomes (function, blood pressure, leg strength, pain, and health-related quality of life) in older Americans with osteoarthritis of the knee and high blood pressure. The modified STAR program addresses the barriers to physical activity from osteoarthritis of the knee as well as high blood pressure–related physical activity concerns. The modified STAR program has the potential to reduce the risk of heart disease in the 5 million older adults who

have both osteoarthritis of the knee and high blood pressure and who do not engage in the recommended amount of physical activity.

Example of a Funded Qualitative Training (F31) Project

Maureen Metzger, a doctoral student at the University of Rochester, submitted a successful application for a NRSA predoctoral (F31) fellowship. The project was funded by NINR in March 2010 and is scheduled to end in March 2012. She prepared the following abstract for a descriptive qualitative study, which was titled "Patients' Perceptions of the Role of Palliative Care in Late-Stage Heart Failure":

Project Summary: Cardiovascular (CV) disease is the leading cause of death in the US, with heart failure (HF) accounting for the majority of deaths from CV disease. Heart failure, which affects more than 5 million people in the US, is a life-limiting condition associated with markedly decreased function and quality of life and high mortality rates. The National Institutes of Health have indicated that a more thorough understanding of the experiences of people confronting life-limiting conditions, including those with non-cancer diagnoses, is warranted. There is consensus that communication with healthcare providers, specifically about prognosis and treatment decisions, is not well managed in late-stage HF, and this is associated with adverse consequences. Many clinicians and researchers have recently been advocating for an increased role of palliative care (PC) consultation in HF and there has been a subsequent trend toward increased referrals to PC services for patients with HF, for goals of care discussions. Despite this trend, the perspectives of HF patients and their family members of PC remain unknown. We do not know what patients and families expect from PC consultations, what their experience of these consultations is, and their perceptions of whether and how PC goals of care discussions affect their treatment planning and decision-making. The proposed qualitative descriptive study will describe the perspectives of 25 HF patient-family member dyads. The specific aims include: 1) To describe the experience of patients with later stage HF and their family members referred to an acute care based PC consultation service for goals of care; and 2) To articulate patients' and family members' perceptions of the role of PC in the care of the patient's disease. Increasing our understanding of the experiences of HF patients and their family members referred for PC consultations would add substantively to the existing body of knowledge in PC and inform the development of future interventions. PUBLIC HEALTH RELEVANCE: Heart failure is a life-limiting and debilitating condition affecting a large number of people in this country. In an attempt to improve the care of patients with later-stage HF, clinicians have been calling for an expanded role of PC in HF. However, in order to design and implement interventions that will appropriately serve patients with HF and the people who love them, we need a better understanding of the experience of HF patients and their family members referred for PC consultations.

SUMMARY POINTS

- A **research proposal** is a written document specifying what a researcher intends to study; proposals are written by students seeking approval for dissertations and theses and by researchers seeking financial or institutional support. The set of skills associated with developing proposals that can be funded is referred to as **grantsmanship**.
- Preparing proposals for qualitative studies is especially challenging because methodologic decisions are made in the field; qualitative proposals need to persuade reviewers that the proposed study is important and a good risk.
- Students preparing a proposal for a dissertation or thesis need to work closely with a well-chosen committee and adviser. Dissertation proposals are often "mini-dissertations" that include sections that can be incorporated into the dissertation.
- The federal government is the largest source of research funds for health researchers in the United States. In addition to regular grant programs through **Parent Announcements**, federal agencies such as the National Institutes of Health (NIH) announce special opportunities in the form of **Program Announcements (PAs)** and **Requests for Applications** (**RFA**s) for **grants** and **Requests for Proposals** (**RFP**s) for **contracts**.

- Nurses can apply for a variety of grants from NIH, the most common being **Research Project Grants** (R01 grants), **AREA Grants** (R15), **Small Grants** (R03), or **Exploratory/Developmental Grants** (R21). NIH also awards training fellowships through the National Research Service Award (NRSA) program as F-series awards and Career Development Awards (K-series awards).
- Grant applications to NIH are submitted online using the SF424, which has a series of special forms (fillable screens) that require uploaded PDF attachments.
- The heart of an NIH grant application is the **research plan component**, which includes two major sections: Specific Aims and Research Strategy. The latter, which is restricted to 12 pages for R01 applications and 6 pages for training fellowships, includes subsections called Significance, Innovation, and Approach.
- NIH grant applications also require a budget, which can be an abbreviated **modular budget** if requested funds for R01 grants do not exceed $250,000 in direct costs per year.
- Grant applications to NIH are reviewed three times a year in a dual review process. The first phase involves a review by a peer review panel (or **study section**) that evaluates each proposal's scientific merit; the second phase is a review by an Advisory Council.
- In NIH's review procedure, the study section assigns **priority** (**impact**) **scores** only to applications judged to be in the top half of proposals based on a preliminary appraisal by assigned reviewers. A score of 10 is the most meritorious ranking, and a score of 90 is the lowest possible score.
- All applicants for NIH grants are sent a summary statement, which offers a critique of the proposal. Applicants of scored proposals also receive information on the priority score and percentile ranking.
- Some suggestions for writing a strong proposal include several for the planning stage (e.g., starting early, selecting an important topic, learning about the audience, reviewing a successful proposal, and creating a strong team) and several for the writing stage (building a persuasive case, justifying methodologic decisions, beginning and ending with a flourish, adhering to proposal instructions, and having the draft proposal critiqued by reviewers).

STUDY ACTIVITIES

Chapter 29 of the *Resource Manual for Nursing Research: Generating and Assessing Evidence for Nursing Practice, 9th ed.*, offers various exercises and study suggestions for reinforcing the concepts taught in this chapter. In addition, the following study questions can be addressed:

1. Suppose that you were planning to study the self-care behaviors of aging AIDS patients.
   a. Outline the methods you would recommend adopting.
   b. Develop a project timeline.
2. Suppose you were interested in studying separation anxiety in hospitalized children. Using references cited in this chapter, identify potential funding sources for your project.

STUDIES CITED IN CHAPTER 29

All references cited in this chapter can be found in a separate section at the end of the book.