Measurement in Nursing and Health Research

Carolyn Feher Waltz
Ora Lea Strickland
Elizabeth R. Lenz
Measurement in Nursing and Health Research
Carolyn Feher Waltz, PhD, RN, FAAN, is a professor at the University of Maryland School of Nursing in Baltimore, Maryland. She received her BSN and MS degrees in public health nursing from the University of Maryland and her PhD in education with specialization in research, evaluation, and measurement from the University of Delaware. Dr. Waltz is an internationally recognized expert in the development of outcomes evaluation and measurement in clinical and educational settings and an adjunct member of the Quebec Interuniversity Nursing Intervention Research Group (GRIISIQ), a consortium of researchers from all Canadian universities in Quebec. She has published extensively, including award-winning books and articles on research, measurement, nursing outcomes, program development, and evaluation. She has received research funding from multiple sources; authored numerous articles, books, and book chapters; and was a founding editor of the journal Evaluation & the Health Professions. She has provided numerous clinical and educational outcomes evaluations, consultations, presentations, and workshops to varied health care audiences nationally and internationally and is widely acknowledged as a pioneer in outcomes measurement, evaluation, and research.

Ora Lea Strickland, PhD, RN, FAAN, is the dean of, and a professor in, the Nicole Wertheim College of Nursing and Health Sciences at Florida International University, Miami, Florida. She has focused her research on women’s and family health, vulnerable populations, and measurement of health-related variables and nursing outcomes. Dr. Strickland is frequently called upon as a consultant nationally and internationally, and has presented more than 300 lectures, speeches, and workshops. She has studied the impact of health and economic policies on the well-being of women and children in southern Africa during her tenure as a Kellogg National Leadership Fellow. Dr. Strickland assisted the National Institutes of Health in the design of the landmark Women’s Health Initiative (WHI) study and was a coprincipal investigator at the Emory site for this study of 168,000 postmenopausal women at 40 national sites over the course of 10 years. Dr. Strickland has served as principal investigator on several other major federally funded programs and NIH grants and conducted research focused on symptoms in expectant fathers, neurometric indices of perimenstrual symptoms, genetic markers of coronary heart disease in premenopausal African American women, African American breast cancer survivors, and emotional outcomes in sickle cell disease. Dr. Strickland served as project director for the Measurement of Nursing Outcomes Project, which she conceptualized with Carolyn Waltz; this program assisted over 200 nurse researchers in the development of nursing outcome instruments. In addition to over 100 refereed journal articles, Dr. Strickland has published 12 books, contributed to 23 other books, and has won nine American Journal of Nursing Book of the Year Awards. She is the founder of the Journal of Nursing Measurement and served as its senior editor for 20 years.
Elizabeth R. Lenz, PhD, RN, FAAN, is a professor emeritus at The Ohio State University College of Nursing, Columbus, Ohio, where she served as dean from 2001 to 2011. Before joining the Ohio State faculty, she held faculty and administrative positions at Columbia University, the Pennsylvania State University, Penn State's Milton S. Hershey Medical Center, the University of Maryland–Baltimore, Georgetown University, and Boston College. She is a leader in doctoral nursing education, having served as the founding director of the PhD program in nursing at the University of Maryland, and is chair of the American Association of Colleges of Nursing’s Taskforce on the Practice Doctorate. Dr. Lenz’s clinical expertise is in community health nursing, and her research has focused on the family as an important context for health and illness and on the quality and outcomes of nurse practitioner primary care practices. She is also known for her theory-building work as an author of the middle-range theory of unpleasant symptoms. Between 1999 and 2001, she served as an editor of the journal Scholarly Inquiry for Nursing Practice.
## Contents

**Contributors**  xi

**Preface**  xiii

**Acknowledgments**  xv

### I  Basic Principles of Measurement  1

1  Introduction to Measurement  3

2  Operationalizing Nursing Concepts  33

3  Measurement Theories and Frameworks  59

*Rosemarie DiMauro Satyshur*

### II  Understanding Measurement Design  111

4  Strategies for Designing Measurement Tools and Procedures  113

*Rosemarie DiMauro Satyshur*

5  Measurement Reliability  183

6  Validity of Measures  209

7  Standardized Approaches to Measurement  261

### III  Measurement via the Digital World  291

8  Internet Data Collection  293

9  Observational Methods  309
Contents

IV Instrumentation and Data-Collection Methods 319

10 Content Analysis 321

11 Interviews 333

12 Questionnaires 349

13 The Delphi Technique 361

14 Visual Analog Scales 371

15 Measurement of Physiological Variables Using Biomedical Instrumentation 379
   Kathleen S. Stone and Susan K. Frazier

16 Precision, Accuracy, and Uncertainty of Biophysical Measurements for Research and Practice 427
   Nancy A. Ryan-Wenger

17 Selection and Use of Existing Instruments 447

V Team Science 457

18 Interprofessional Practice Measurement and Evaluation 459
   Susan Antol

19 Metrics and Benchmarking in Quality Improvement in Health Professions Education and Practice 467
   With Margaret Scisney-Matlock
VI Measurement Issues 485

20 Ethical Issues 487

21 Measurement Issues in Qualitative Research 507
   Judith E. Hupcey and Lisa Kitko

22 Measurement Issues in Translational Science 529
   Linda L. Chlan and Mary Beth Happ

23 Other Measurement Issues 539

Index 597
Contributors

Susan Antol, PhD, RN
Assistant Professor
Partnerships, Professional Education, and Practice
University of Maryland School of Nursing
Baltimore, Maryland

Linda L. Chlan, PhD, RN, FAAN
Associate Dean for Nursing Research
Nursing Research Division
Department of Nursing
Mayo Clinic
Rochester, Minnesota

Susan K. Frazier, PhD, RN, FAHA
Associate Professor
Director, PhD Program
Co-Director, RICH Heart Program
University of Kentucky
College of Nursing
Lexington, Kentucky

Mary Beth Happ, PhD, RN, FGSA, FAAN
Nursing Distinguished Professor and
Associate Dean for Research and Innovation
The Ohio State University
College of Nursing
Columbus, Ohio

Judith E. Hupcey, EdD, CRNP, FAAN
Professor of Nursing and Medicine
College of Nursing
College of Medicine
The Pennsylvania State University
Hershey, Pennsylvania
Contributors

Lisa Kitko, PhD, RN, CCRN
Assistant Professor of Nursing
College of Nursing
The Pennsylvania State University
University Park, Pennsylvania

Nancy A. Ryan-Wenger, PhD, RN, CPNP, FAAN
Nurse Scientist
Nationwide Children’s Hospital
Investigator, Center for Innovation in Pediatric Practice
The Research Institute at Nationwide Children’s Hospital
Columbus, Ohio

Rosemarie DiMauro Satyshur, PhD, RN
Assistant Professor
University of Maryland
School of Nursing
Baltimore, Maryland

Margaret Scisney-Matlock, PhD, RN, FAAN
Professor, Nicole Werttheim College of Nursing and Health Sciences
Florida International University
Miami, Florida
Professor Emeritus
University of Michigan–Ann Arbor
Ann Arbor, Michigan

Kathleen S. Stone, PhD, RN, FAAN
Professor Emeritus
The Ohio State University
College of Nursing
Columbus, Ohio
Preface

The fifth edition of this text was written taking into account the needs and interests of nurses and their expanded audience of health researchers, the significant increase in interdisciplinary research collaboration, and growing emphasis on evidence-based practice as an important research outcome across all health disciplines. The intent, within the context of these trends, is to present a pragmatic account of the process involved in designing, testing, selecting, and/or evaluating instruments, methods, and procedures for measuring variables in clinical, educational, and research settings.

We continue to strive to meet the needs of a large and diversified audience that ranges from neophytes to those with more advanced knowledge and experience in measurement. Therefore, we do not assume that most readers have an extensive background and experience in measurement or statistics. Rather, we begin our discussion of content assuming little background in these areas and subsequently develop, explain in detail, and illustrate by example the concepts and principles operationally important to the content presented. In this manner, it is possible for the less sophisticated reader to develop the level of knowledge necessary to understand the content that is included for the benefit of the more advanced reader. This book should serve as a valuable resource for readers who seek basic and advanced content to develop their skill in measurement.

This edition includes the “best” of the measurement content, strategies, and procedures presented in the previous editions and provides the most up-to-date content, strategies, and procedures available that have direct applicability for nurses and health researchers in a variety of roles, including those of student, educator, clinician, researcher, administrator, and consultant. As in previous editions, the focus is on increasing the reader’s ability to employ measures that are operationalized within the context of theories and conceptual frameworks, derived from sound measurement principles and practices, and adequately tested for reliability and validity using appropriate methods and procedures. Throughout this edition, examples and studies conducted by nurses and health researchers in varied settings and/or in collaboration with others in their own and other disciplines are provided to illustrate important content and to reinforce the importance of using sound measurement principles and practices within the context of evidence-based practice. Attention is given to measurement issues resulting from changes in nursing and health research that have increased in number and complexity since the time of the last edition. Additional reference sources, which are readily available in libraries and/or online, are provided for readers who desire to pursue further the topics presented. Whenever possible, comprehensive summaries of literature in an area of interest are cited rather than a myriad of individual books and articles.
Some key features of this edition:

- Chapter 6 provides updated and expanded reconceptualizations, definitions, and strategies for testing the validity of measures.
- Chapter 8 contains expanded content on issues of data privacy and how to address them.
- The expansion of Chapter 15 focuses on the methods for measuring physiological variables using biomedical instrumentation and issues to be addressed, including use in translational research.
- A new Chapter 16 presents content regarding the application of measurement principles and practices in clinical research and practice that focuses on clinical data-collection methods, including clinimetrics, and the unique challenges of conducting research in clinical settings and situations in which strict experimental control and standardization are not optimal and how to address this; standards for evaluating precision and accuracy of clinical measures; and issues to be addressed relative to use of outcomes as a basis for clinical practice.
- The content of Chapter 21 has been expanded to include additional material relevant to the use of a mixed-methods approach to enhance measurement outcomes.

We appreciate the positive feedback we have received from satisfied readers who have valued previous editions as textbooks for research and measurement courses and from those conducting their own research and/or seeking to evaluate the quality of others’ research to decide whether or not to use their results as a basis for practice. Also, we thank the many of you who requested another edition and sincerely hope that you will find this fifth edition a significant resource.

Carolyn Feher Waltz
Ora Lea Strickland
Elizabeth R. Lenz
Acknowledgments

The authors sincerely appreciate the contributions of Sandra A. Murphy, whose commitment, perseverance, organizational ability, and many talents were instrumental to us in preparing this manuscript. Sandy, we can’t thank you enough for so graciously contributing your time and effort to ensure the completion of this book. You are the best!
Basic Principles of Measurement
Introduction to Measurement

The use of sound measurement principles and practices is an essential component of well-designed research studies, which is of upmost importance when the goal is to employ research results as a basis for practice. In this chapter, terms and ideas essential to understanding the content of subsequent chapters provide an overview of the types of measures most often used and introduce the two essential characteristics of any measuring tool or method: reliability and validity.

THEORY-BASED MEASURES

A theory is defined by Polit and Beck (2012, p. 744) as an abstract generalization that presents a systematic explanation about the relationships among phenomena. Kerlinger (1973, p. 9) defined a theory as “a set of interrelated constructs, definitions and propositions that present a systematic view of phenomena by specifying relations among variables, with the purpose of explaining and predicting the phenomena.” Polit and Beck (2012) note that a theory “is a systematic, abstract explanation of some aspect of reality. In a theory, concepts are knitted together into a coherent system to describe or explain some aspect of the world” (p. 57). A theoretical rationale, according to LoBiondo-Wood and Haber (1994), “provides a road map or context for examining problems, and developing and testing hypotheses, gives meaning to the problem and study findings by summarizing existing knowledge in the field of inquiry and identifying linkages among concepts” (p. 157). Conceptual models provide a context for constructing theories that deal with phenomena of concern to a discipline and help define how it is different from other disciplines. Polit and Beck (2012) use the terms conceptual model and conceptual framework interchangeably. In their view a conceptual model includes “interrelated concepts or abstractions assembled together in a rational scheme by virtue of their relevance to a common theme” (p. 749).

For example, Arwood and Kaakinen (2009) designed a simulation model based on language and learning that was evolved from Arwood’s Neuro Sematic Language Learning Theory that provides a hierarchical framework for assessing and measuring conceptual learning outcomes. Evans et al. (2009) employed the biopsychosocial model to understand the health benefits of yoga. They present and review a conceptual model of the potential biopsychosocial benefits of yoga and empirical evidence that supports the involvement of physical, psychological, and spiritual domains as possible mechanisms of actions of yoga upon well-being. Krogh and Naden (2008) report on a project they undertook using an information management system to
Basic Principles of Measurement

develop a conceptual framework that would provide clinicians with an approach to
documentation consistent with legal and organizational requirements and enable
them to retain the ability to record all aspects of clinical nursing. The result was the
Norwegian documentation KPO model, which incorporates the Nursing Minimum
Data Set and NANDA (nursing diagnosis, nursing intervention classification, and
nursing outcome classification) and focuses on quality assurance, problem solving,
and caring.

Various authors have defined and analyzed concepts in a number of ways
(Allegood & Marriner-Tomey, 1997; George, 2002; Glaser & Strauss, 1967; Leonard
& George, 1995; Meleis, 1997; Orem, 1995; Polit & Beck, 2012; Renpenning & Taylor,
2003). Simply defined, a concept is a thought, notion, or idea. It is an abstraction.
For example, nursing concepts are thoughts, notions, or ideas about nursing or nurs-
ing practice. Thus, concepts define the content of interest in measuring phenom-
ena. Phenomena are observable facts or events. To render concepts measurable it is
necessary to translate them into phenomena. When one operationalizes a concept,
one translates an abstract concept into concrete observable events or phenomena.

For example, the concept of “attitude” is frequently operationalized as a tendency to
respond in a consistent manner to a certain category of stimuli (Campbell, 1963). If
the stimulus is a 17-item questionnaire regarding children's concerns about health
care, such as the Child Medical Fear Scale developed by Broome and Mobley (2003,
pp. 196–206), in which the child indicates whether he or she was “not at all,” “a little,”
or “a lot” afraid of selected experiences associated with health care described in the
items, and the subject responds “a lot” to the majority of the 17 items, one would infer
from the responses that the child's attitude or fearfulness of the experiences associated
with health care was high.

Variables are quantities that may assume different values; they are changeable. The
process whereby one decides how to measure a variable is referred to as instrumentation,
that is, the process of selecting or developing tools and methods appropriate for measur-
ing an attribute or characteristic of interest. In the previous example, the 17-item ques-
tionnaire was a form of instrumentation selected to measure attitudes of children toward
the medical experience. Instrumentation is a component of the measurement process.
Measurement is defined as the process of assigning numbers to objects to represent the
kind and/or amount of attributes or characteristics possessed by those objects. This defi-
nition of measurement includes what has traditionally been referred to as qualitative
measurement (i.e., assigning objects to categories that represent the kind of characteristic
possessed and that are mutually exclusive and exhaustive) as well as quantitative mea-
surement (i.e., assigning objects to categories that represent the amount of a characteristic
possessed).

The utilization of a conceptual framework to systematically guide the measure-
ment process increases the likelihood that concepts and variables universally salient
to nursing and health care practice will be identified and explicated. That is, when
measurement concerns emanate from an empirical rather than a conceptual point of
view, there is higher probability of investigating and measuring these variables from an
esoteric or limited perspective that overlooks important dimensions of the variables
that should be measured. Concepts of interest to nurses and other health professionals are usually difficult to operationalize, that is, to render measurable. This is explained in part by the fact that nurses and other health professionals deal with a multiplicity of complex variables in diverse settings, employing a myriad of roles as they collaborate with a variety of others to attain their own and others’ goals. Hence, the dilemma that they are apt to encounter in measuring concepts is twofold: first, the significant variables to be measured must somehow be isolated; and second, very ambiguous and abstract notions must be reduced to a set of concrete behavioral indicators. What tools, therefore, are available to nurses and other health care professionals who must begin to grapple with this dilemma?

Because of the increased interest in evidence-based practice (EBP) across health care disciplines and the challenge to provide services of broadening scope and diversity in order to keep pace with the rapidly changing and volatile health care scene coupled with severe shortages in nursing and other health care disciplines, a great deal of controversy and confusion has ensued regarding which functions should be included within the realm of each of the practice disciplines and which information should be shared. For nursing, this is evidenced by the proliferation of definitions and models for nursing practice evident in the literature. Although, for the most part, the definitions of nursing advanced in the literature remain ambiguous and global, in each view the major focus for nursing practice can be placed on a continuum ranging from direct to indirect involvement in patient care. Direct nursing practice involves the continuous, ongoing provision of direct services to patients and clients (e.g., the primary care nurse practitioner provides direct nursing services). Indirect nursing practice is usually characterized by activities on behalf of the patient; that is, working with or through others who are directly responsible for the provision of direct services to patients and clients. Nursing education, administration, and health policy activities exemplify indirect nursing practice. This scheme for categorizing nursing practice has utility for the nurse who is attempting to operationalize nursing concepts.

More specific, the first task is to identify a definition of nursing that is consistent with nurses’ own views and beliefs about nursing practice. Similarly, although the extent to which available conceptual frameworks and models for nursing practice have been refined and tested varies, their very existence affords nurses a rich opportunity to select a conceptual framework to guide them in systematically identifying and further explicating concepts and variables germane to nursing and nursing practice concerns within their primary focus. The problems, settings, roles, and purposeful activities undertaken by nurses will differ, depending upon whether their primary focus is direct or indirect nursing practice. Furthermore, the goals for, and outcomes likely to result from, the application of direct and indirect nursing processes will vary. Although there will be differences among each of these categories of practice, there will also be commonalities in process and outcomes within each focus. Therefore, if nurses consider their measurement concerns within the context of their primary focus, delimit the processes and outcomes that characterize that practice, and then search for behavioral indicators within the primary focus that extend beyond their immediate setting (i.e., that are common across settings similar to their own), they are
apt to reduce the abstract concepts emanating from their conceptual framework to
behavioral indicators with more universal acceptance than those likely to result from
a more esoteric approach. In this manner, nurses will ultimately make a contribution
to the profession by accruing information to add to the body of knowledge about
nursing, the specific definition and conceptual framework employed, and its utility
as a guide for operationalizing nursing and nursing practice concerns. It should be
noted that when nurses whose measurement concerns emanate from their ongoing
practice fail to step back and rethink the problem from a conceptual point of view,
they also have a high probability of investigating and measuring their variables from a
limited perspective that overlooks important dimensions of the variables that should
be measured.

For example, Stacey et al. (2009) report on their efforts to integrate a patient deci-
sion support theoretical framework and associated evidence-based resources through-
out a baccalaureate nursing curriculum. Rew, Grady, Whittaker, and Bowman (2008)
employed social cognitive theory as a basis for their study to determine the effects of
duration of homelessness and gender on personal and social resources, cognitive–
perceptual factors, and sexual health behaviors among the homeless. The work of
Ellenbecker, Porell, Samia, Byleckie, and Milburn (2008) exemplifies testing of a the-
oretical model of home health care nurse retention. The use of conceptual models as a
basis for measurement is illustrated by the work of Landis, Parker, and Dunbar (2009)
who employ a biobehavioral–ecological framework in their study of sleep, hunger, sati-
ety, food cravings, and caloric intake in adolescents, and Avci and Kurt (2008), who
employ the Health Belief Model in their study of health beliefs and mammography
rates of Turkish women living in rural areas. A conceptual approach within a qualitative
phenomenological study was undertaken by Melby, Dodgson, and Tarrant (2008), who
sought to describe the experiences of Western expatriate nursing educators teaching in
eastern Asia. Additional examples of theories, conceptual frameworks, and theory-based
measures can be found in Bramadat and Drieger (1993); Kempen, Miedema, Ormel, and
Molenar (1996); Fry (2001); Armitage (2001); Weber, Kopelman, and Messick (2004);
March and Olsen (2006); Kaji, Koenig, and Bey (2006); Olsen (2007).

In nursing and the health professions, there is often a salient concern with the
measurement of process variables, such as with developing and implementing process
measures of quality, which are dynamic, as well as outcome variables, which are usually
static, and in which results of the measurement are likely to be applied to the solu-
tion of significant problems across practice settings as in EBP. For example, Rubin,
Pronovost, and Diette (2001) recognize the importance of rigorously developing
quality indicators when developing process measures “that will provide insights into
opportunities to improve quality of care” and identify the following steps to be taken.
“Developing a process measure includes defining the purpose of and audiences for
the measures, choosing the clinical area to evaluate, organizing the assessment team,
choosing the component of the process to measure, writing the indicator specifi-
cations, performing preliminary tests of feasibility, reliability and validity, and determin-
ing scoring and analytical specifications” (pp. 489–496). The measurement of outcome
variables is exemplified in the work of Schneider, Barkauskas, and Keenan (2008) who
conducted a study to determine the sensitivity and responsiveness of the Outcome and Assessment Information Set (OASIS) and the Nursing Outcomes Classification (NOC) to the effects of home health care nursing intervention (p. 76).

Two salient characteristics of all measurement efforts are reliability and validity. First and foremost, tools and methods selected or developed for measuring a variable of interest must demonstrate evidence for reliability and validity. Reliability refers to the consistency with which a tool or method assigns scores to subjects. Validity refers to the determination of whether or not a tool or method is useful for the purpose for which it is intended, that is, it measures what it purports to measure. Second, in addition to the concern with instrument reliability and validity, attention needs to be given to the reliability and validity of the measurement process per se. To increase the probability that the measurement process will yield reliable and valid information, it is necessary whenever possible to employ multiple tools or methods to measure any given variable (all of which have demonstrated evidence for instrument reliability and validity) and to obtain information about any given variable from a number of different perspectives or sources.

For example, Aitken et al. (2008) conducted a study “to describe strategies used to administer an international multicenter trial to assess the effectiveness of a nursing educational intervention” (p. 101). Noting the importance of clinical relevance and that “in order to support appropriate development of an evidence base for practice, nursing interventions should be tested in multiple settings,” they employed multiple sites, countries, and methods (Aitken et al., 2008, p. 101). Lee, Chaboyer, and Wallis (2008) in their study to predict health-related quality of life 3 months after traumatic injury, employed multiple measurement methods, including telephone interviews, clinical data, and questionnaires using rating scales.

Measurement reliability and validity are thus largely functions of a well-designed and well-executed measurement process. For this reason, the intent of this book is to provide the reader with a sound background in the theories, principles, and practices of measurement and instrumentation that are germane to the measurement of concepts in nursing and the health professions.

MEASUREMENT FRAMEWORKS

Just as it is important to identify and employ a conceptual framework for determining what is to be measured and delineating how it will be operationalized, it is equally important to identify and employ a measurement framework to guide the design and interpretation of the measurement per se. The two major frameworks for measurement are the norm-referenced and the criterion-referenced approaches.

A norm-referenced approach is employed when the interest is in evaluating the performance of a subject relative to the performance of other subjects in some well-defined comparison or norm group. The Stress of Discharge Assessment Tool (SDAT-2) developed by Toth (2003, pp. 99–109) is an example of a 60-item norm-referenced measure of the stress experienced by patients at hospital discharge and in the early recovery period at home following acute myocardial infarction.
Scores on each item in the SDAT-2 range from 1 to 5 points, depending on the patient's degree of agreement with the item. A high score indicates high stress for that item, and a low score indicates low stress. The total possible score ranges from 60 to 300 points, and its value for a given subject takes on meaning when it is considered in light of the scores obtained by other patients who responded to the same tool.

Similarly, the results of the application of physiologic measures, such as blood pressure readings, are often interpreted on the basis of readings (usually ranges of values) considered normal for some well-defined comparison group (e.g., Black males older than 40 years of age with no significant health problems). It should be noted that the terms “norm referenced” and “standardized” are not synonymous. Standardized tests are one type of norm-referenced measure; there are other types as well. Unlike most other norm-referenced measures, a standardized measure is designed by experts for wide use and has prescribed content, well-defined procedures for administration and scoring, and established norms. The Graduate Record Examination (Stein & Green, 1970), the National League for Nursing Achievement Test Battery (Waltz, 1988), and nurse practitioner certification examinations, such as the Neonatal Intensive Care Nursing Examination (Perez-Woods, Burns, & Chase, 1989), are examples of standardized measures.

A key feature of a norm-referenced measure is variance. The task when using a norm-referenced measure is to construct a tool or method that measures a specific characteristic in such a way that it maximally discriminates among subjects possessing differing amounts of that characteristic, that is, spreads people out along the possible ranges of scores. For example, if the characteristic to be measured is knowledge of human sexuality content, then test items are designed to differentiate among individuals with varying levels of knowledge of the content. The goal is to obtain scores in such a manner that the result is a few high scores, with most scores in the middle range and a few low scores. If this goal is achieved, the resulting distribution of scores on the measure should look much like a normal curve. Figure 1.1(A) illustrates the distribution of scores that one would expect to result from the employment of the hypothetical five-item, norm-referenced measure of human sexuality content.

The sole purpose of a criterion-referenced measure is to determine whether or not a subject has acquired a predetermined set of target behaviors. The task in this case is to specify the important target behaviors precisely and to construct a test or measure that discriminates between those subjects who have and those who have not acquired the target behaviors. How well a subject's performance compares with the performance of others is irrelevant when a criterion-referenced framework is employed. Criterion-referenced measures are particularly useful in the clinical area when the concern is with the measurement of process and outcome variables.

For example, a criterion-referenced measure of process would require that one identify standards for the patient care intervention and then compare subjects’ clinical performance with the standards for performance (i.e., predetermined target behaviors) rather than compare subjects’ performance with that of other subjects, all of whom may not meet the standards. Standards might be those resulting from international and/or national professional organizations, such as Guidelines for the Management of Patients.
With Chronic Stable Angina, published by the American College of Cardiology and the American Heart Association (2007), or local site-specific standards such as those agreed upon by health care providers serving on the EBP committee in a local oncology outpatient setting regarding the management of chemotherapy-induced nausea and vomiting in their setting. Similarly, when a criterion-referenced measure is employed to measure patient outcomes, a given patient’s status is determined by comparing his or her performance with a set of predetermined criteria (e.g., EKG normal, diastolic pressure below 80, other vital signs stable for 4 hours post-op) or target behaviors (e.g., requests for pain medication have ceased by 2 days post-op, desire to return to normal activities is verbalized by third day post-op) rather than by comparing his or her performance with that of other patients.

**FIGURE 1.1** (A) Normal distribution of scores on a five-item norm-referenced measure. (B) Positively skewed distribution of scores on a five-item criterion-referenced measure.
Basic Principles of Measurement

King’s (2003, pp. 3–20) Measure of Goal Attainment Tool is an example of a criterion-referenced measure of functional abilities and goal attainment behaviors. This tool was constructed to assess individuals’ physical abilities to perform activities of daily living and the behavioral response of individuals to the performance of these activities. Goal attainment was defined as mutual goal setting by nurse and patient, and assessment of goal attainment. Each of these areas comprises three components: (a) personal hygiene, (b) movement, and (c) human interaction. Essential tasks in performing actions related to each of these areas are reflected in items contained on the tool. Percentage scores are determined to evaluate the independence or dependence of subjects related to the essential tasks. Thus, the individual’s performance, as reflected by the score, is interpreted on the basis of his or her ability to perform the essential tasks and is not related to the evaluation of the performance of others using the same tool.

One would expect the distribution of scores resulting from a five-item, criterion-referenced measure to look like the illustration in Figure 1.1(B). It should be noted that not only does the distribution of scores resulting from a criterion-referenced measure have less variance or spread than that resulting from a norm-referenced measure, but it also is skewed in shape. In a skewed distribution, scores tend to cluster at one end of the scale; in the example in Figure 1.1(B), the cluster is at the high end. A more detailed discussion of score spread (variance) and distribution shape (normal and skewed) is presented in Chapter 3.

Because the design, scoring, interpretation, reliability, and validity testing for norm-referenced and criterion-referenced measures differ, it is important to decide which of the two measurement frameworks will be employed prior to the conduct of any other steps in the measurement process.

TYPES OF MEASURES

Once the conceptual basis for the measure and the measurement framework has been determined, attention turns to the selection of the specific type of measure to be employed. Semantics is often a problem for newcomers to the field of measurement. Many different terms are employed to label like measures. For this reason, it is important to consider some of the classification schemes and terms that are used to label the different types of measures. In addition to being categorized as either norm-referenced or criterion-referenced, approaches to measurement are also usually referred to as qualitative or quantitative in nature. Whether the approach is qualitative, quantitative, or mixed methods will largely be a function of the measurement problem, the conceptual basis guiding the determination of what is to be measured, and the nature of the variables to be measured.

When a qualitative approach to measurement is employed, objects are assigned to mutually exclusive and exhaustive categories that represent the kind of characteristic possessed by the object. Qualitative measurement methods are usually indicated when the theoretical orientation for the measurement derives from phenomenology, existentialism, symbolic interactionism, or ethnomethodology (Chenitz & Swanson, 1986, p. 3). Knafl (1989) suggests that within nursing, ethnography, phenomenology,
and grounded theory appear to be the dominant approaches. The major goal of qualitative methods is to document and interpret as fully as possible the whole of what is being measured from the frame of reference of the subjects involved (Diekelmann, 1992; Duffy, 1987; Filstead, 1974; Girot, 1993). Hinds and Young (1987) note that qualitative methods usually attempt to describe processes and thus tend to be measures of dynamic and changing phenomena. For example, Turris (2009) employed a grounded theory approach in a study of women’s decisions to seek treatment for the symptoms of potential cardiac illness. Data for their study were obtained from in-depth interviews with women who went to one of two emergency departments within a specific time period for the treatment of symptoms indicating potential cardiac arrest. Hendrickson (2008) conducted a study to explore worries, safety behaviors, and perceived difficulties in keeping children safe at home in a sample of low-income predominately non-English-speaking mothers as a foundation for later nursing intervention. They employed a qualitative descriptive design with content analysis of responses to three semistructured interview questions (p. 137). A phenomenological approach was employed by Melby et al. (2008) to describe the lived experience of English-speaking Western nurse educators teaching in East Asian countries. Subjects were interviewed about their experiences and resulting narrative data were analyzed by each member of the research team “to identify themes and then through group discussions to reach consensus, develop clear understandings of language nuances and maintain the participant’s voices” (p. 176). Mbweza, Norr, and McElmurry (2008) used a grounded theory approach to examine “the decision-making processes of husband and wife dyads in matrilinear and patrilinear marriage traditions of Malawi in the areas of money, food, pregnancy, contraception and sexual relations” (p. 12).

Although the measurement strategies and techniques employed across the qualitative theoretical perspectives may differ somewhat, the types of measures employed are often the same and are also employed when quantitative approaches are used. Qualitative measurement methodologies generally include content analysis of documents; reviews of the literature and of findings from studies to identify common themes; participant and nonparticipant observations; interviews or focus groups that may be structured or nonstructured, but are usually open-ended; and open-ended questionnaires. Qualitative data collection and analysis procedures are discussed in Chapter 21.

Quantitative measurement assigns objects to categories that represent the amount of a characteristic possessed by the object. Quantitative methods emphasize the search for facts and causes of human behavior through objective, observable, and quantifiable data (Duffy, 1987). Hinds and Young (1987) suggest that quantitative approaches provide outcome data and information on the representativeness of the studied sample and thus tend to be measures of more stable phenomena. Specific types of methods employed with quantitative approaches include the variety of types of measures discussed in later sections of this chapter.

Single studies of a problem using qualitative or quantitative methods rarely involve full exploration of the problem area (Bergstrom, 1989). When a mixed-methods approach, that is, when a combination of both qualitative and quantitative methods are employed together, the two approaches to measurement provide information
12  I  Basic Principles of Measurement

regarding the internal and external validity of the studies or measurement processes (Campbell & Fiske, 1959; Vedich & Shapiro, 1955; Webb, Campbell, Schwartz, & Sechrest, 1966). For example, Escoffery, Glanz, Hall, and Elliott (2009) employed a multimethod process evaluation for a skin cancer prevention diffusion trial. Data-collection methods in their study included surveys, database tracking, activity logs, process evaluation interviews and site visits. Thus, to develop an adequate and useful repertoire of measurement principles and practices, one needs to understand when and how to use both qualitative and quantitative approaches to measurement, as well as how they can be combined as measurement methodologies.

Stainbeck and Stainbeck (1984) note that qualitative and quantitative approaches, because they derive from different perspectives, have several inherent differences that should be understood prior to considering how the two methods can complement each other in the measurement of nursing phenomena. Major differences noted by them include the following: When quantitative approaches are employed, the goal is to arrive at an understanding of a phenomenon from the outsider’s perspective by maintaining a detached, objective view that hypothetically is unbiased. The perspective, on the other hand, when qualitative approaches are employed is that of an insider and the goal is to obtain information by talking to and/or observing subjects who have experienced firsthand the phenomena under scrutiny.

1. Quantitative methods focus on the accumulation of facts and causes of behavior assuming that facts gathered do not change, whereas qualitative methods are concerned with the changing, dynamic nature of reality.

2. When quantitative approaches are used, the situation is structured by identifying and isolating specific variables for measurement and by employing specific measurement tools and methods to collect information on these variables. In contrast, qualitative approaches attempt to gain a complete or holistic view of what is being measured by using a wide array of data, including documents, records, photographs, observations, interviews, case histories, and even quantitative data.

3. Usually highly structured procedures, designed to verify or disprove predetermined hypotheses, are employed with quantitative approaches. Flexibility is kept to a minimum in an attempt to minimize bias. Procedures used with qualitative approaches, on the other hand, are usually flexible, exploratory, and discovery oriented.

4. Quantitative approaches yield objective data that are typically expressed in numbers, whereas qualitative approaches focus on subjective data that are typically expressed or reported through language.

5. Quantitative data are usually collected under controlled conditions, whereas qualitative data are usually collected within the context of their natural occurrence.

6. In both approaches reliability and validity are valued. In the quantitative approach, there is a heavy emphasis on reliability, that is, data that are consistent, stable,
and replicable. Qualitative approaches, while recognizing that reliability is a necessary prerequisite for validity, tend to concentrate on validity, that is, data that are representative of a true and full picture of the phenomenon that is investigated (pp. 130–131).

Over time, more researchers have come to value using a mixed-methods approach and have begun to recognize the value of integrating qualitative and quantitative approaches within the context of a given study. It should be noted that the integration of qualitative and quantitative approaches is not simply mixing methods, but rather requires one to assume that the two approaches are complementary and that the primacy of the paradigmatic assumptions underlying one or the other approach can be eliminated as unproductive (Haase & Myers, 1988). Triangulation, discussed in more detail in Chapter 20, is one methodological strategy for combining qualitative and quantitative approaches. In triangulation, multiple data sources, collection techniques, theories, and investigators are employed to assess the phenomenon of interest (Fielding & Fielding, 1986; Madey, 1982; Mitchell, 1986). Examples of the use of triangulation for combining qualitative and quantitative approaches can be found in Breitnauer, Ayres, and Knafl (1993); Floyd (1993); Corey (1993); S. A. Mason (1993); and Hendrickson (2008).

In addition to being categorized as norm-referenced or criterion-referenced, qualitative or quantitative, measuring tools and methods may be classified by (a) what they seek to measure, (b) the manner in which responses are obtained and scored, (c) the type of subject performance they seek to measure, or (d) who constructs them.

WHAT IS MEASURED?

In nursing and health care research, there is usually interest in measuring cognition, affect, psychomotor skills, and/or physical functioning. Cognitive measures assess the subject's knowledge or achievement in a specific content area. Indicators of cognitive behavior usually are obtained as follows:

1. Achievement tests (objective and essay) that measure the extent to which cognitive objectives have been attained.
2. Self-evaluation measures designed to determine subjects' perceptions of the extent to which cognitive objectives have been met.
3. Rating scales and checklists for judging the specific attributes of products produced in conjunction with or as a result of an experience.
4. Sentence-completion exercises designed to categorize the types of responses and enumerate their frequencies relative to specific criteria.
5. Interviews to determine the frequencies and levels of satisfactory responses to formal and informal questions raised in a face-to-face setting.
6. Peer utilization surveys to ascertain the frequency of selection or assignment to leadership or resource roles.
7. Questionnaires employed to determine the frequency of responses to items in an objective format or number of responses to categorized dimensions developed from the content analysis of answers to open-ended questions.
14  ■  Basic Principles of Measurement

8. Anecdotal records and critical incidents to ascertain the frequency of behaviors judged to be highly desirable or undesirable.

9. Review of records, reports, and other written materials (e.g., articles, autobiographical data, awards, citations, honors) to determine the numbers and types of accomplishments of subjects.

The number of cognitive measures employed far exceeds the number of other types of measures. Specifically, written multiple-choice tests are the most often used, perhaps because they are the most objective of the various cognitive measures and the most reliable, and because they have the greatest utility in measuring all types of knowledge. Multiple-choice tests are further discussed in Chapters 4 and 14. Examples of cognitive measures can be found in Smith (1991); Grant et al. (1999); Story (2001); Arnold (2001); Tiro, Meissner, Korbin, and Chollete (2007); Boon, Nelson, Laufman, Khort, and Kozinetz (2007); and Rondahl (2009); and design of mail and Internet surveys are discussed in Dillman (2007). It should be noted that cognitive measures are not limited to paper-and-pencil tests. A variety of other approaches exist, including Internet data collection discussed in Chapter 8.

Affective measures seek to determine interests, values, and attitudes. Interests are conceptualized as preferences for particular activities. Examples of statements relating to interests are:

- I prefer community-based nursing practice to practice in the hospital setting.
- I like to work with student nurses as they give care to patients.
- I prefer teaching responsibilities to administrative responsibilities.
- I would enjoy having one day a week to devote to giving direct care to patients in addition to my teaching responsibilities.

Values concern preferences for life goals and ways of life, in contrast to interests, which concern preferences for particular activities. Examples of statements relating to values are:

- I consider it important to have people respect nursing as a profession.
- A nurse's duty to her patient comes before duty to the community.
- Service to others is more important to me than personal ambition.
- I would rather be a teacher than an administrator.

Attitudes concern feelings about particular social objects, that is, physical objects, types of people, particular persons, or social institutions. Examples of statements relating to attitudes are:

- Nursing as a profession is a constructive force in determining health policy today.
- Continuing education for nurses should be mandatory for relicensing.
- Humanistic care is a right of all patients.
- All nurses should be patient advocates.
The feature that distinguishes attitudes from interests and values is that attitudes always concern a particular target or object. In contrast, interests and values concern numerous activities: specific activities in measures of interest and very broad categories of activities in measures of value. It is extremely difficult to preserve the conceptual differences among interests, values, and attitudes when actually constructing measures of affect. Thus, for the purpose of rendering them measurable, they are all subsumed under the rubric of acquired behavioral dispositions (Campbell, 1963) and are defined as tendencies to respond in a consistent manner to a certain category of stimuli. For example, when patients are asked to respond to a questionnaire to indicate their satisfaction with the quality of care received, one is interested in measuring their tendency to consistently respond that they are satisfied or dissatisfied, given a set of questions that ask them about the care they received (the stimuli). Examples of the use of affective measures can be found in Grice, Picton, and Deakin (2003); Denny-Smith, Bairan, and Page (2005); Mackler, Wilkerson, and Cinti (2007); and Gerend and Maglorie (2008).

Self-report measures are the most direct approach to the determination of affect. In this type of measure subjects are asked directly what their attitudes, interests, or values are. For example, subjects might be given a list of favorable and unfavorable statements regarding antagonistic patients and asked to agree or disagree with each. Such a self-report inventory is referred to as an attitude scale. Other indicators of affective behaviors include, but are not limited to:

1. Sentence-completion exercises designed to obtain ratings of the psychological appropriateness of an individual's responses relative to specific criteria
2. Interviews
3. Questionnaires
4. Semantic differential, Q-sort, and other self-concept perception devices
5. Physiologic measures
6. Projective techniques, for example, role playing or picture interpretation
7. Observational techniques and behavioral tests, including measures of congruence between what is reported and how an individual actually behaves in a specific situation
8. Anecdotal records and critical incidents

Examples of self-report measures are included in Mackler, Wilkerson, and Cinti (2007) and Gerend and Maglorie (2008).

From the empirical evidence concerning the validity of different approaches, it appears that self-report offers the most valid approach currently available. For this reason, at present, most measures of affect are based on self-report and usually employ one of two types of scales: a summated rating scale or a semantic differential scale. A scale is a measuring tool or method composed of:

1. A stem, which is a statement relating to attitudes or an attitudinal object to be rated by the respondent
2. A series of scale steps
3. Anchors that define the scale steps
I. Basic Principles of Measurement

An example of a scale can be found in Chiravalle and McCaffrey (2005). Figure 1.2 presents examples of the components of a scale.

There are different types of anchors that can be employed: numbers, percentages, degrees of agreement/disagreement, adjectives (e.g., worthless/valuable), actual behavior, and products (e.g., samples of nursing care plans to be rated 1–6). Usually numerical anchors are preferred because if the meaning of each step on the scale is specified at the beginning of the rating form, as is usually the case, numbers provide an effective means of coordinating those definitions with rating scales; numbers on scales constantly remind subjects of the meanings of scale steps; and numbers facilitate the analysis of data, for example, inputting ratings for computer analysis (Nunnally, 1967; Nunnally & Bernstein, 1994).

**SUMMATED RATING SCALE**

A summated rating scale contains a set of scales, all of which are considered approximately equal in attitude or value loading. The subjects respond with varying degrees of intensity on a scale ranging between extremes such as agree/disagree, like/dislike, or accept/reject. The scores of all scales in the set are summed or summed and averaged to yield an individual's attitude score. An example of a summated rating scale is given in Figure 1.3.

Summated rating scales are easy to construct, are usually reliable, and are flexible in that they may be adapted for the measurement of many different kinds of attitudes. Nunnally (1967) and Nunnally and Bernstein (1994) suggest that the reliability of summated scales is a direct function of the number of items. When there are a reasonable number of items (e.g., 20) on the scale, fewer scale steps for individual scales are required for a high degree of reliability. When there are fewer items, more scale steps for individual scales are required for reliability. In most cases, 10 to 15 items using five or six steps are sufficient. Individual scales on summated attitude scales tend to
correlate substantially with each other, because it is fairly easy for the constructor to devise items that obviously relate to each other and for subjects to see the common core of meaning in the items. Additional information regarding summated attitude scales can be found in Edwards (1957), Shaw and Wright (1967), Nunnally (1967), and Nunnally and Bernstein (1994).

An example of a summated rating scale can be found in Lehoux, Richard, Pineault, and Saint-Arnaud (2006).

**SEMANTIC DIFFERENTIAL SCALE**

The semantic differential is a method for measuring the meaning of concepts that was developed by Osgood, Suci, and Tannenbaum (1957). The semantic differential has three components: (a) the concept to be rated in terms of its attitudinal properties, (b) bipolar adjectives that anchor the scale, and (c) a series of five to nine scale steps (seven is the optimal number of steps suggested). Figure 1.4 presents an example of a semantic differential scale. The concept to be rated in Figure 1.4 is “noncomplying patient.” Respondents are instructed to rate the concept according to how they perceive it or feel about it by placing an X along the 7-point scale anchored by the bipolar adjective pairs. The resulting scale response can be converted to numerical values and treated statistically.

Nunnally (1967) and Nunnally and Bernstein (1994) explain that the logic underlying the semantic differential stems from the recognition that, in spoken and written language, characteristics of ideas and objects are communicated largely by adjectives. It is reasonable on this basis to assume that meaning often can be and usually is communicated by adjectives; it is also reasonable to assume that adjectives can be used to measure various facets of meaning. The semantic differential primarily measures...
connotative aspects of meaning, that is, what implications the object in question has for the respondents. For example, if an individual rating the concept “noncomplying patient,” said, “I dislike them very much,” this statement would represent a connotation or sentiment for that type of patient. The semantic differential is one of the most valid measures available for assessing the connotative aspects of meaning, particularly the evaluative connotations of objects. An example of a semantic differential scale can be found in Rempusheski and O’Hara (2005).

Factor analytic studies of semantic differential scales have suggested that there are three major factors of meaning assessed by such scales: (a) evaluation, (b) potency, and (c) activity. Table 1.1 presents the pairs of adjectives most frequently used to define each of these factors. Additional information regarding semantic differential scales can be obtained from Osgood et al. (1957) and Snider and Osgood (1969). In Chapter 14, issues related to the selection and use of two additional types of scales that are being given increased attention by nurses and other health professionals—visual analog and magnitude estimation—are discussed. Additional examples of affective measures and scales can be found in the work of Cousins (1997), Czar and Engler (1997), Adam and Freeman (2001), Kelly (2001), and Chiravalle and McCaffrey (2005).

Psychomotor measures seek to assess subjects’ skills, that is, their ability to perform specific tasks or carry out specific procedures, techniques, and the like. An important consideration in the measurement of psychomotor objectives involves the manner in which the skills and materials or objects to be manipulated or coordinated are specified. Specifically, criteria for the successful manipulation of an object must be clearly and unambiguously stated at the time when objectives are made explicit. Task-analysis procedures (Gagne, 1962) are often used to accomplish this.

The most viable approach to the measurement of psychomotor skills at this time is the observation method combined with a performance checklist or rating scale. The observation method always involves some interaction between subject and observer in which the observer has an opportunity to watch the subject perform.

**FIGURE 1.4 Example of a semantic differential scale.**

Rate the following concept in terms of how you feel about it at this point in time:

Noncomplying patient

<table>
<thead>
<tr>
<th>/</th>
<th>/</th>
<th>/</th>
<th>/</th>
<th>/</th>
<th>/</th>
<th>/</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

Ineffective

Effective

<table>
<thead>
<tr>
<th>/</th>
<th>/</th>
<th>/</th>
<th>/</th>
<th>/</th>
<th>/</th>
<th>/</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

Weak

Strong
Checklists and rating scales are often employed to record in a systematic manner behaviors and events occurring during the observation period. A checklist is most often used to note the presence or absence or the frequency of occurrence of the specified behavior, conditions, characteristics, or events. Rating scales are employed when it is desirable that the observer rate the behavior in terms of points along a continuum. Ratings may occur at predetermined intervals throughout the observation session or at the end of the observation to summarize what occurred during the observation period.

The conceptual definition of the phenomenon to be measured forms the basis for how and where the observation should occur. Qualitative conceptual perspectives generally lead to free and unstructured observations in naturalistic settings, whereas quantitative conceptualizations more often lead to more structured observations using guidelines and trained observers. The work of Cohn, Matias, Tronick, Connell, and Lyons (1986) provides an example of a study combining both structured, standardized observation methods and naturalistic unstructured observations of depressed mothers and their infants. The reader will find an interesting comparison and contrast between the two approaches for collecting observational data.

Unstructured and semistructured observations involve the collection of descriptive information that is generally analyzed in a qualitative manner. When structured observation is employed, it is often necessary that the nurse prepare an observation guide to structure the observation and train the observer in its use. This guide increases the probability that the crucial behaviors of concern will be considered, which increases the reliability and validity of the method.

**TABLE 1.1 Frequently Employed Anchors for Semantic Differential Factors**

<table>
<thead>
<tr>
<th>Evaluation</th>
<th>Potency</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bipolar Adjectives</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Good</td>
<td>Strong</td>
<td>Active</td>
</tr>
<tr>
<td>Bad</td>
<td>Weak</td>
<td>Passive</td>
</tr>
<tr>
<td>Fair</td>
<td>Large</td>
<td>Quick</td>
</tr>
<tr>
<td>Unfair</td>
<td>Small</td>
<td>Slow</td>
</tr>
<tr>
<td>Positive</td>
<td>Severe</td>
<td>Tense</td>
</tr>
<tr>
<td>Negative</td>
<td>Lenient</td>
<td>Relaxed</td>
</tr>
<tr>
<td>Honest</td>
<td>Hard</td>
<td>Sharp</td>
</tr>
<tr>
<td>Dishonest</td>
<td>Soft</td>
<td>Dull</td>
</tr>
<tr>
<td>Successful</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unsuccessful</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Valuable</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Worthless</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
20

Basic Principles of Measurement

When possible, the guide should specify when an observation begins and ends as well as what behaviors are to be observed. Frequently, time is the vehicle for accomplishing this purpose. For example, it might be specified that observations of medical students’ history taking begin when the student enters the room and continue for the first 5 minutes of the student–patient interaction. Using a more elaborate scheme, observations begin when the student enters the room, and will continue for 2 minutes; then observers will rest for the next 2 minutes, rate for 2 minutes, rest for 2 minutes, and so forth, until the student leaves the room and the encounter ends. No matter how structured the observation or how well trained or competent the observer, observation techniques, in order to be sound, require more than one observer. This provides an estimate of the accuracy or reliability of the observations and provides a basis for determining the degree of confidence to be placed in the data.

Three factors must be considered in the discussion of observational techniques: (a) interaction between respondents and observers, (b) whether or not respondents know they are being observed, and (c) whether or not respondents know when they will be observed. Observation is difficult because watching a situation often changes it so that the observers are no longer certain of what they are observing. This implies that a basic criterion for evaluating studies in which observation is used is the extent to which the situation observed was natural. Observations of a subject’s psychomotor skills should be accomplished with as little effect as possible on the natural situation in which the skills are normally performed. Webb et al. (1966) have published a useful book full of suggestions about how measures can be collected as unobtrusively as possible.

It is necessary to weigh the value of collecting observational data over time as opposed to collecting information at one isolated point in time. Observational data collected at one time are subject to more errors of measurement and, hence, lower reliability and validity than observational data collected at multiple times. When subjects’ psychomotor performance is of interest, the concern is usually with how they perform most of the time or typically; that is, patterns of performance or consistency in performance over time becomes important. When observational data are collected at one point in time, there is greater probability that the results of the measurement will reflect more of the conditions surrounding that isolated point in time than the true abilities of the subjects to perform the tasks or behaviors. Hence, whenever possible, measures of performance should occur at more than one point in time.

Observational techniques may be direct or indirect. In direct observation the observer evaluates psychomotor performance by simply watching the subject perform. A limitation of this approach stems from the fact that it is both time consuming and expensive. It is, however, an excellent technique for the assessment of behavior in conjunction with clinical performance, especially when the concern is with dynamic or process variables. Similarly, a unique strength of the observation method results from the fact that if the observer wishes to learn how a subject functions under the pressure of supervision, there is no substitute for direct observation that is known and scheduled.
Indirect observation methods include motion picture, television, videotaping, and other devices for recording subjects’ activities. The value of indirect techniques results from the opportunities they afford subjects to become involved in the evaluation of their performance as the recording is viewed jointly by the observer and respondents. Indirect observations are limited in that they are not sensitive to the tone, mood, or effect of the situation. Another limitation is that mechanical devices selectively record, depending on their placement, where they are aimed by the operator and, hence, the total situation may be missed. This limitation can be turned into an advantage, however, if multiple devices are used to record all that happens before them. Examples of psychomotor measures can be found in Bujak, McMillan, Dwyer, and Hazleton (1991); DeMattes et al. (1993); D. J. Mason and Redeker (1993); Finke et al. (2001); Mims (2001); Kostopoulos (2001); Gilbert, Temby, and Rogers (2004); Philpin (2006); Williams (2006); and Cricco-Lizza (2006). Observational methods are discussed in more detail in Chapter 9.

Physiologic measures seek to quantify the level of functioning of living beings. Indicators of physiologic functioning include but are not limited to:

1. Blood pressure readings
2. Temperature readings
3. Respiratory measures
4. Metabolic readings
5. Diabetic and other screening devices
6. Readings from cardiac and other monitoring instruments
7. EKG and EEG readings
8. Results of blood tests and analyses
9. Measures of height and weight

Physical functioning can often be measured by a scientific instrument, and the results of physiologic measures usually are expressed as a quantitative scale that can be graded into finely distinguished numerical values. For example, the variable diastolic blood pressure is measured using a scientific instrument referred to as a sphygmomanometer. Its scale is in a quantitative form ranging from 0 to 300, providing a total of 300 different continuous scale points or values to which a subject can be assigned and which differentiate among the various degrees of the variable possessed by the subjects measured. Thus, on the basis of blood pressure readings, one can state that a subject with a diastolic pressure of 100 is 20 points higher than one with a diastolic pressure of 80. This 20-point difference is significant in comparing the physical status of two patients.

Well-designed and implemented physiologic measures are among the most precise methods one can employ; they yield data measured at the interval or ratio level of measurement, allow a wide range of statistical procedures to be employed in their analysis, and tend to produce results that demonstrate a high degree of reliability and validity. Examples of physiologic measures can be found in Heidenreich and Giuffre (1990), Bridges and Woods (1993), and Partridge and Hughes (2007). The physiologic approach to measurement is discussed in detail in Chapter 12.
HOW RESPONSES ARE OBTAINED AND SCORED

The distinction to be considered is whether a measure is objective or subjective. It should be noted that a given method or technique is generally viewed as more or less objective or subjective; that is, one may think in terms of a continuum anchored by the terms objective and subjective, and then place a given method on the continuum, depending on whether it possesses characteristics more like those of an objective or subjective measure.

**Objective measures** contain items that allow subjects little if any latitude in constructing their responses and spell out criteria for scoring so clearly that scores can be assigned either by individuals who know nothing of the content or by mechanical means. Multiple-choice questions and physiologic measures are examples of the most objective methods that can be employed.

**Subjective measures** allow respondents considerable latitude in constructing their responses. In addition, the probability that different scorers may apply different criteria is greater. Examples of subjective measures are the essay test, open-ended interview questions, case studies, and nursing care plans. The essay question is a method requiring a response constructed by the subject, usually in the form of one or more sentences. The nature of the response is such that (a) no single answer or pattern of answers can be listed as correct, and (b) the quality of the response can be judged only subjectively by one skilled or informed in the subject (Stalnaker, 1951). Thus, significant features of the essay method are (a) the freedom of response allowed the respondents, (b) the fact that no single answer can be identified as correct or complete, and (c) responses must be scored by experts who themselves usually cannot classify a response as categorically right or wrong. Essay questions may require subjects to express their own thoughts on an issue of interest to the profession, outline a research design for investigating a research question, derive a mathematical proof, or explain the nature of some nursing phenomenon. Items may require only a brief response or may demand an extensive exposition.

Advocates of the essay approach argue that an important characteristic of individuals is their ability to interact effectively with other individuals in the realm of ideas. The basic tool of interaction is language, and successful individuals are those who can react appropriately to questions or problems in their field as they encounter them. It is not enough, they contend, to be able to recognize a correct fact when it is presented or to discriminate among alternatives posed by others. Successful individuals are the masters of their collection of ideas and are able to cite evidence to support a position and contribute to the advancement of ideas and constructs within their field. The only way to assess the extent to which individuals have mastered a field is to present them with questions or problems in the field and assess how they perform. Hence, they argue, the essay format provides an avenue for assessing scholarly and/or professional performances better than other available methods (Coffman, 1971).

Even so, because of their subjective nature, essay questions have inherent limitations that must be recognized and minimized if sound measurement is to result from
their use. The limitations of essays and other subjective measures fall into two general categories: (a) problems related to the difficulty in achieving consistency in scoring responses and (b) problems associated with the sampling of content. Empirical evidence regarding the reliability of subjective measures suggests that different raters tend to assign different scores to the same response, a single rater tends to assign different scores to the same response on different occasions, and the differences tend to increase as the measure permits greater freedom of response (Finlayson, 1951; Hartog & Rhodes, 1936; Noyes, 1963; Pearson, 1955; Vernon & Millican, 1954). Different raters may differ as a result of a number of factors, including the severity of their standards, the extent to which they distribute scores throughout the score scale, and real differences in the criteria they are applying.

Basic to problems associated with the sampling of content is the notion that each sample unit should be independent and equally likely to be chosen in the sample. In general, the greater the number of different questions, the higher the reliability of the score. The compromise to be made, however, is between the desire to increase the adequacy of the sample of content by asking many different questions, and the desire to ask questions that probe deeply into the subjects’ understanding. Additional information regarding essay items are presented in Stalnaker (1951) and Coffman (1971). Examples of structured, unstructured, and semistructured interviews can be found in Irwin and Johnson (2005), Cricco-Lizza (2006), and Clark et al. (2006), respectively. Other subjective measures are described in more detail in Part IV.

**TYPE OF PERFORMANCE MEASURED**

When performance is of interest, one may seek to measure typical performance or maximum performance. If the interest is in assessing subjects as they do their best (produce their highest quality work), then a *maximum performance* measure is appropriate. Such measures are indices of cognition that generally measure a set of skills a subject possesses but that differ among themselves in the specificity of their focus and the use to which scores are put. Maximum performance measures of particular interest include aptitude measures, achievement measures, and diagnostic measures.

*Aptitude* tests are specific measures of capacity for success and tend to focus on various general aspects of human ability (e.g., mechanical aptitude, artistic aptitude). They are often used as predictors of performance in special fields.

*Achievement* measures are tests of particular skills and knowledge and are more specific than aptitude tests. They usually sample a wide range of skills and are constructed by nurses and other health professionals for their own use. Commercially produced achievement measures are also available in many different content areas. *Diagnostic* tests are even more specific in their focus than achievement measures, although this need not always be the case. They focus on specific skills and often employ multiple measures of particular skills. Their intent is to pinpoint specific weaknesses that might not be apparent otherwise. Once specific deficiencies are identified and remediation has taken place, one might predict that achievement, which is assumed to be dependent on these more specific skills, will improve.
24  I  ■  Basic Principles of Measurement

If information about subjects’ typical behaviors (i.e., what they usually do or would do) is of interest, it is appropriate to use a typical performance measure. These are measures of affective behavior and usually attempt to have respondents describe the way they typically perceive themselves or their behavior. Typical performance measures usually ask the subjects for scaled responses, forced-choice responses, or criterion-keyed responses. Figure 1.5 presents examples of each of these types of responses.

### Scaled response

When a scaled response is employed, the respondent indicates on a scale what his or her rating or answer is to a question posed. For example:

Do findings from this research study provide information that will be meaningful to you in your clinical practice? Please rate.

<table>
<thead>
<tr>
<th>not at all</th>
<th>very little</th>
<th>somewhat</th>
<th>enough</th>
<th>a lot</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

### Forced-choice response

With a forced-choice response item, the respondent is asked to choose between two or more different alternatives, all of which may be appealing responses. The point is that one particular response is most appealing to the subject. For example:

A program of ongoing evaluation and implementation of research findings that may serve as a basis for practice does not exist in the agency with which you are affiliated. You are aware of the need to develop such a program. You would prefer to have this need met by:

1. Referring the task to someone else.
2. Supporting activities of the professional nursing organizations that are seeking to increase involvement in implementing evidence-based practice in all clinical settings.
3. Supporting a policy change in the agency responsible for care.
4. Serving as a resource person to staff by providing them with knowledge and materials to enable them to develop such a program.
5. Becoming a member of a committee of practitioners who are developing and testing a pilot program in conjunction with the patients for whom they deliver care.
6. Initiating the idea of such a program by implementing evidence-based practice for patients on your unit and sharing with staff the various approaches you employ in evaluating and implementing research findings in your practice.

### Criterion-keyed response

Criterion-keyed responses depend on information previously obtained about how certain groups answered the items. If a subject's score looks like those of members of a predefined group, he or she is classified as a member of that group. The assumption is that the criterion for membership in a specific group is having a set of responses on the measure that looks like those from the predefined group. For example the Minnesota Multiphasic Personality Inventory (MMPI) was originally used with hospitalized psychiatric patients and normal (i.e., nonhospitalized subjects) to construct a criterion-keyed set of questions that had some value as predictors of mental stability, that is, if a specific item was responded to differently by the two groups, it was included on the measure.

**FIGURE 1.5** Sample responses of typical performance measures.
WHO CONSTRUCTS MEASURES?

*Standardized* measures are developed by specialists for wide use. Their content is set, the directions for administration (often including time limits) are clearly described, and the scoring procedure to be used is completely prescribed. Information on norms concerning scores is generally available. Examples of standardized measures employed to assess the outcomes of nursing education programs are presented in *Educational Outcomes: Assessment of Quality—A Prototype for Student Outcome Measurement* (Waltz, 1988). In Chapter 7, standardized approaches to measurement are discussed in detail.

*Informal tools and methods* are typically constructed by nurses and other health professionals for their own use. They are not content constrained; that is, the user is free to define the content as well as administration procedures and scoring. Norms may be available for local groups but more often are not available for any group.

In summary, the measurement framework employed in a given situation will have important implications for instrument development and for what can be done with and on the basis of the resulting information. Thus, it is important to clarify at the outset the type of measurement that will yield data appropriate for the types of questions and/or hypotheses one seeks to answer. In Chapters 2 through 19, attention is focused on instrument development and testing in both the norm-referenced and criterion-referenced cases. In Chapters 20 to 23, measurement issues and important considerations to be made in using the types of measures presented in this section are addressed.

RELIABILITY AND VALIDITY OF MEASURES

As indicated in the foregoing sections reliability and validity are essential characteristics of any measuring tool or method. Factors that may affect the degree of consistency obtained for a given measure (reliability) are (a) the manner in which the measure is scored, (b) characteristics of the measure itself, (c) the physical and/or emotional state of the individual at measurement time, and (d) properties of the situation in which the measure is administered (e.g., the amount of noise, lighting conditions, temperature of the room).

Strictly speaking, one validates not the measurement tool or method but rather some use to which the measure is put. For example, an instrument designed to select participants who would benefit from a primary care fellowship experience must be valid for that purpose, but it would not necessarily be valid for other purposes such as measuring how well participants master objectives at the completion of the fellowship experience.

Both reliability and validity are matters of degree rather than all-or-none properties. Measures should be assessed each time they are used to see if they are behaving as planned. New evidence may suggest modifications in an existing measure or the development of a new and better approach to measuring the attribute in question. Reliability is a necessary prerequisite for validity; that is, if a measure does not
assign scores consistently, it cannot be useful for the purpose for which it is intended. Reliability is not, however, a sufficient condition for validity; that is, because a measure consistently measures a phenomenon does not ensure that it measures the phenomenon of interest.

As stated earlier, the determination of the reliability and validity of a specific tool or method will differ depending on whether it is norm referenced or criterion referenced. Specific techniques for determining reliability and validity in each case are discussed in Chapters 5 and 6. In either case, the reliability and validity of the measurement process itself is increased when multiple measures of the same thing are employed; that is, more than one type of instrumentation is used to answer a given question. Similarly, reliability and validity increase when the answer to a given measurement concern is elicited by collecting data from a number of different sources using the same measurement tool or method.

Evidence for reliability and validity of a tool or method is accrued over time, so although one cannot rely on “old” reliability and validity evidence, it is important to seek information regarding reliability and validity testing results in previous studies. Often difficulties are encountered in locating information regarding reliability and validity evidence for existing tools. An article by Strickland (2006) presents strategies for searching for evidence of reliability and validity for an existing instrument that may be useful in this regard.

REFERENCES


Validating Nursing Diagnosis. Palm Springs, CA: Cosponsored by the North American Nursing Diagnosis Association and Case Western University.


1 • Introduction to Measurement

30 • Basic Principles of Measurement


Measurement Reliability

The use of sound measurement practices and procedures must be a salient concern during the design, implementation, and evaluation of a research study, especially in regard to assessing the quality of research results for use as a basis for practice (Polit & Beck, 2012). Thus, pretesting the measures and data-collection methods employed in a research study is an essential component that enables the researcher to assess the appropriateness of the type of data-collection method employed, identify needed revisions, and evaluate reliability and validity. Rubin, Pronovost, and Diette (2001) in their article regarding the development and testing of quality indicators, emphasize the dangers of not pretesting when using hospital records as a sole source for measuring whether left ventricular function has been assessed as a measure for congestive heart failure. Specifically, they note that in some settings this assessment is usually performed by an outpatient physician and transmitted to an attending physician in the hospital and thus is often missing from the inpatient medical record (p. 494).

A pretest is a trial run of a measure and data-collection methods that are undertaken to provide information regarding the method's reliability and validity and to reveal problems relating to its content, administration, and scoring. The measure must be pretested, usually on a small sample of subjects for whom it was designed, under conditions that approximate as nearly as possible the conditions expected to exist when it is employed. For example, Kip, Ehlers, and van der Wal (2009) in a study of patients' adherence to antiretroviral therapy in Botswana pretested their face-to- face structured interview using a number of approaches. Specifically, (a) each item on the interview schedule was judged by two researchers and four HIV/AIDS clinicians to ascertain whether it measured the theoretical construct it purported to measure; (b) Cronbach's alpha coefficient was employed to determine whether there were evidencethat items within each subscale were internally consistent; (c) administration procedures and potential problems with items were assessed by interviewing 10 patients who were excluded from participation in the actual study; and (d) 12 faculty members, two experienced nurse researchers, and a statistician reviewed the interview schedule for content and construct validity. Another example can be found in the work of Glaister (2007) who, in a study to determine how the presence of mathematical and computer anxiety in nursing students affects learning of dosage calculations, pretested her dosage competency instrument with a group of six students who had successfully completed studies in the content area the preceding semester.

During the conduct of the pretest, it is important to be attentive to the reactions, comments, and nonverbal communication of respondents that might give clues...
to problems with the measure. Similarly, observations and concerns during the administration that may suggest needed improvements should be recorded. For example, problems related to maintain interest, questions raised by respondents, adequacy of the time provided to respond, test length, and the like may come to light during the pretest. It is also most beneficial after the pretest data have been collected to ask respondents to identify difficulties they have encountered in completing the measure, suggestions they may have for improving it, and possible discrepancies between the purpose for which items were constructed and how subjects understood and responded to the items. In the earlier study, Glaister (2007) conducted focus groups after completion of the dosage competency instrument with eight randomly selected participants from each intervention group in her study.

In addition, scores should be computed and data compiled and tabulated for interpretation, including the preparation of tables and graphs, so that any difficulties with scoring, interpretation, or preparation of the data for analysis will be evident. Appropriate procedures for estimating the method’s reliability and validity should be employed, including item-analysis procedures using the pretest data. On the basis of the information obtained from the pretest, especially the resulting evidence for the measure’s reliability and validity, a decision should be made concerning whether the method will be used as is or needs modification before it can be employed. If it is determined that the method needs modifications for improvement, these should be made and the method pretested again prior to its use.

When sufficient evidence for reliability and validity is obtained as a result of the pretest, the measure may then be employed for data collection, but its reliability and validity should still be monitored each time it is employed, using less extensive and more economical procedures than required for the pretest. Should monitoring of the measure suggest that reliability and/or validity are not holding up with use, it is necessary to scrutinize the measure using more rigorous and extensive reliability and validity studies to ascertain needed changes, to make the modifications required, and then to pretest again (Waltz & Bausell, 1981). Another example of pretesting can be found in the work of Blood-Siegfried et al. (2008) who developed an evaluation rubric to measure quality in a graduate online curriculum and conducted a pretest to assess its utility and applicability in five online, core, master’s level–courses that resulted in revision.

Norm-referenced measures are derived from classical measurement theory. In Chapter 3, it was noted that in this view, every observed score \(O\) is composed of a true score \(T\), which represents the precise amount of the attribute possessed by the subject at measurement time, and an error score \(E\). If a large number of subjects are measured on the attribute in question and their observed scores plotted, reliability would be conceptualized as the proportion of the variance in the observed score distribution that is due to true differences in subjects’ possession of the attribute being measured. Unreliability would be conceptualized as the proportion of variance in the observed score distribution that is due to error. Hence, in this view every measurement involves some error that, although it can never be eliminated in total, can be reduced.
Measurement error may be random or systematic. If the nurse had only one thermometer and it was accurate, but she misread it while obtaining different measures, the error would be random. Random errors limit the degree of precision in estimating the true scores from observed scores, and therefore, lead to ambiguous measurement and decreased reliability of the measure. In practice, reliability concerns the extent to which measurements are repeatable by the same individual using different measures of the same attribute or by different individuals using the same measure of an attribute. Thus, research and evaluation efforts are limited by the reliability of measuring instruments and/or reliability with which they are employed. More specifically, sources of random error include, but are not limited to, imprecision in the measure itself, temporal factors, individual differences at measurement time, and/or imprecision in the administration or scoring of the measure.

If, in the preceding example, the nurse employed the thermometer correctly, but the thermometer itself was inaccurate and always registered 0.5 points higher than it should, the error in the nurse’s measurement would be systematic. This systematic or constant error would contribute to the mean score of all subjects equally and thus would become part of the true score of each individual. Because validity is defined as the extent to which an instrument measures what it purports to measure, systematic errors, because they affect the true scores of all subjects, would decrease the validity of the measure rather than its reliability.

In Chapter 3, it was noted that reliability is a necessary but not sufficient condition for validity; that is, a measure that demonstrates evidence for reliability will not necessarily demonstrate evidence for validity as well. The amount of random error places a limit on measurement validity, but even in the complete absence of random errors there is no guarantee of measurement validity; that is, the correlations between a tool and an independent criterion can never be higher than the square root of the product of the reliability of the two and the reliability of the criterion variable (Issac & Michael, 1995). Similarly, because random error may occur as a result of circumstances surrounding the administration of the measure and/or individual differences at measurement time, reliability, and validity investigations conducted on one measurement occasion are not sufficient evidence for reliability and validity when measures are employed on other occasions or with different subjects. Thus, evidence for reliability and validity must be determined every time a given measure is employed.

CONCEPTUAL BASIS FOR RELIABILITY

The determination of reliability in the norm-referenced case is conceptualized using the domain-sampling model. As noted in Chapter 3, this model views any particular measure as composed of a random sample of items from a hypothetical domain of items. For example, an adjective checklist designed to measure anxiety in presurgical patients would be thought of as containing a random sample of adjectives from all possible adjectives reflective of anxiety in that patient group. Obviously, the model does not hold strictly true empirically, because it is usually not practical or feasible to explicate all possible items defining a domain of interest, thus items are actually
randomly sampled for a specific measure. The model does, however, lead to principles and procedures for determining evidence for reliability that have much utility in practice.

On the basis of the domain-sampling model, the purpose for any measure is to estimate the measurement that would be obtained if all the items in the domain were employed. The score that any subject would obtain over the whole domain is the subject’s true score. To the extent that any sample of items on a given measure correlates highly with true scores, the sample of items is highly reliable. In other words, specific measures are viewed as randomly parallel tests that are assumed to differ somewhat from true scores in means, standard deviations, and correlations because of random errors in the sampling of items. Thus, in this view, the preferred way to estimate the reliability of a measure is to correlate one measure with a number of other measures from the same domain of content. Because in practice this is often impractical, usually one measure is correlated with only one other measure to obtain an estimate of reliability. The domain-sampling model suggests that the reliability of scores obtained on a sample of items from a domain increases with the number of items sampled. Thus, a one-item measure would be expected to have a small correlation with true scores, a 10-item measure a higher correlation, and a 100-item measure an even higher correlation.

**NORM-REFERENCED RELIABILITY PROCEDURES**

In the norm-referenced case, reliability is usually estimated using a test–retest, parallel form, and/or internal consistency procedure.

The test–retest procedure is appropriate for determining the quality of measures and other methods designed to assess characteristics known to be relatively stable over the time period under investigation. For this reason, test–retest procedures are usually employed for investigating the reliability of affective measures. Because cognitive measures assess characteristics that tend to change rapidly, this procedure is not usually appropriate for estimating their reliability.

When a test–retest procedure is employed, the concern is with the consistency of performance one measure elicits from one group of subjects on two separate measurement occasions. To estimate test–retest reliability for a given measure, one would:

1. Administer the instrument or method under standardized conditions to a single group of subjects, representative of the group for which the measure was designed.
2. Readminister the same instrument or method under the same conditions to the same group of subjects. Usually the second administration occurs approximately 2 weeks after the first, although the time may vary slightly from setting to setting. It should be noted that it is important to ascertain that no activities have occurred between the first and second administration, which may have affected the stability of the characteristic being measured.
3. Determine the extent to which the two sets of scores are correlated. When data are measured at the interval level, the Pearson product–moment correlation coefficient \( r_{xy} \) that is discussed in Chapter 3 is taken as the estimate of reliability. When data are
measured at the nominal or ordinal level, a nonparametric measure of association, such as Chi square-based procedures or Spearman rho, is used. Discussion of the conditions under which specific correlation coefficients are appropriately used, as well as their computation, may be found in Waltz and Bausell (1981), Nunnally and Bernstein (1994), Munro (2005), and Polit and Beck (2012). An example of test–retest reliability can be found in Whiting and Mallory (2007) who conducted a longitudinal study to determine the effects of mentoring on middle school students by nursing and other college students. The correlation coefficient, $r_{xy}$, was employed to compare mean item scores on the Child Behavior Checklist (CBCL) and Teachers Report Form (TRF) on two occasions. Examples of test–retest reliability also can be found in Hall, Rayens, and Peden (2008) and Wang and Chiou (2008).

The value of the reliability coefficient resulting from a test–retest procedure reflects the extent to which the measure rank orders the performances of the subjects the same on the two separate measurement occasions. For this reason, it is often referred to as the coefficient of stability. The closer the coefficient is to 1.00, the more stable the measuring device is presumed to be.

When it is desirable to employ a more stringent index of test–retest reliability, that is, to determine the absolute agreement between the two sets of scores, the percentage of agreement index is calculated. Engstrom (1988) advocates the percentage of agreement as an essential index for describing the reliability of physical measures because it reflects both the precision of measurement and frequency of error and has direct and useful clinical meaning. That is, in most clinical situations, some measurement error is acceptable, but there is a limit on the amount of error that can be tolerated without jeopardizing patient safety. This limit, she notes, can be used as an index of agreement. More specific discussion regarding the assessment of the reliability and validity of physical measures is presented in Chapter 6. For example, Arozullah et al. (2007) employed percentage of agreement in developing and validating a short-form, rapid estimate of adult literacy in medicine for assessing patient literacy in diverse research settings for use when designing interventions for clinical contexts. Chang and Roberts (2008) employed percentage of agreement between the investigator and a nurse with research and clinical experience with older adults concerning the amount of food eaten by residents in an observational study undertaken to investigate factors related to feeding difficulty that is shown in interaction between nursing assistants and elderly residents with dementia.

Whenever two forms of an instrument can be generated, the preferred method for assessing reliability is the parallel form procedure. In parallel form reliability, the interest is in assessing the consistency of performance that alternate forms of a measure elicit from one group of subjects during one administration. Two measures are considered alternate or parallel if they have (a) been constructed using the same objectives and procedures, (b) approximately equal means, (c) equal correlations with a third variable, and (d) equal standard deviations.

Prior to assessing parallel form reliability, it is necessary to obtain empirical evidence that the two measures meet these four criteria. To provide empirical evidence
for equal means and standard deviations, both measures are administered to the same group of subjects on the same occasion, and a test of the significance of the difference between the means and a homogeneity of variance test are employed. If the resulting means are not statistically and significantly different and the variances are homogeneous, evidence that the two measures are parallel is said to exist. Similarly, to obtain empirical evidence that both measures have equal correlations with a third variable, a measure of a third variable believed to be highly correlated with the phenomena being assessed by the parallel measures, which has demonstrated evidence for reliability and validity, is administered to the same group of subjects on the same occasion as the two measures believed to be parallel. Evidence that the two measures are parallel is said to exist if the scores resulting for each of the two measures correlate significantly with the scores resulting from the measurement of the third variable.

Given evidence for parallel forms, to estimate parallel form reliability one would:

1. Administer two alternative forms of a measure to one representative group of subjects on the same occasion or on two separate occasions.
2. Determine the extent to which the two sets of scores are correlated, using an appropriate parametric or nonparametric correlation coefficient as an estimate of reliability.

If both forms of the measure are administered on the same occasion, the value of the resulting reliability coefficient reflects form equivalence only. If the measure is administered on two occasions, stability as well as form equivalence is reflected. Values above 0.80 are usually taken as evidence that the forms may be used interchangeably. An example of the parallel measures model can be found in Finke et al. (2003).

Internal consistency reliability is most frequently employed for cognitive measures where the concern is with the consistency of performance of one group of individuals across the items on a single measure. To estimate the internal consistency of a measure, one would administer the measure under standardized conditions to a representative group on one occasion. The alpha coefficient $KR_{20}$ or $KR_{21}$ would be calculated as the estimate of reliability.

The alpha coefficient is the preferred index of internal consistency reliability because it (a) has a single value for any given set of data, and (b) is equal in value to the mean of the distribution of all possible split-half coefficients associated with a particular set of data. Alpha represents the extent to which performance on any one item on an instrument is a good indicator of performance on any other item in the same instrument. An example of determination of the alpha coefficient follows.

**Example 5.1:** Five newly diagnosed diabetics are given a 10-item multiple-choice test to assess their knowledge and understanding of diabetic food exchanges and the scores in Table 5.1 are obtained. Using the information in Table 5.1:

1. There are 10 items on the tool, the number of items $- 1 = 9$
2. The variance of the test score distribution is 2.64
3. The sum of the variances for each of the 10 items is 2.04 that is $(0.24 + 0.4 + 0.4 + 0.4 + 0.2 + 0.2 + 0 + 0 + 0 + 0.2)$

4. Alpha = \((\text{the number of items/\text{the number of items}} - 1) \times \left[1 - \left(\text{sum of the item variances/\text{variance of the test score distribution}}\right)\right]\)

$$\alpha = \left(\frac{10}{9}\right) \left[1 - \left(\frac{2.04}{2.64}\right)\right]$$

$$\alpha = (1.11)(1 - 0.77)$$

$$\alpha = (1.11)(0.23)$$

$$\alpha = 0.2553,$$ rounded to 0.26

The resulting alpha value of 0.26 indicates the test has a very low degree of internal consistency reliability, that is, the item intercorrelations are low. As a result, performance on any one item is not a good predictor of performance on any other item. A high alpha value is usually taken as evidence that the test as a whole is measuring just one attribute, for example, knowledge of diabetic food exchanges, which in the example is not the case. When tests are designed to measure more than one attribute (e.g., those with subscales or components), alpha should be determined for each scale or subset of homogeneous items in addition to the test as a whole.

A number of factors surrounding the measurement situation may affect the alpha value obtained, and for this reason, it is wise to consider the following when alpha is employed:

1. Alpha is a function of test length. The longer the test, that is, the more items included, the higher the resulting alpha value.

2. A spuriously high alpha may be obtained in a situation in which it is not possible for most respondents to complete the test or measure. As a rule of thumb, Martuza (1977) suggests that if less than 85% of the subjects respond to all items on the test,
alpha should not be used as an estimate of reliability. Equivalently, alpha should not be used when speeded tests are employed.

3. As with all reliability estimates, alpha should be determined each time a test is employed.

4. From the example, it is apparent that alpha is dependent upon the total test variance; that is, the higher the value of the total test variance, the greater the alpha value obtained.

5. Alpha is dependent upon the shape of the resulting distribution of test scores. When a skewed test-score distribution results, variance is usually less than that obtained when the distribution approximates a normal curve, and hence, alpha may be lower in value. Similarly, when alpha is employed with a group of subjects homogeneous in the attribute being measured, alpha will be lower than when a heterogeneous group is measured.

KR 20 and KR 21 are special cases of alpha used when data are dichotomously scored, that is, when each item in a test is scored 1 if correct and 0 if incorrect or missing.

**Example 5.2:** Using the data from Table 5.1, KR 20 is determined as follows:

1. There are 10 items on the tool, the number of items − 1 = 9
2. The variance for item one = the proportion of correct responses to the item (2/5 or 0.4) multiplied by the proportion of incorrect responses (3/5 or 0.5), that is (0.4) (0.5) = 0.24. The variances for the remaining 9 items are 0.24, 0.24, 0.24, 0.16, 0.16, 0, 0, 0, and 0.16, respectively.
3. The sum of the item variances is 1.44.
4. The variance for the test-score distribution is 2.64
5. **KR 20** = (number of items/number of items − 1) multiplied by \[1.00 − (\text{sum of item variances/variance of the test score distribution})\]

\[
KR 20 = \frac{10}{9}[1.00 - (1.44 / 2.64)] \\
= (1.11)(1.00 - 0.54) \\
= (1.11)(0.46) \\
= 0.51
\]

If one can assume that the difficulty level of all items is the same, that is, that the proportion of correct responses is the same for all items, KR 21 may be employed. Because this is not the case for the items in Table 5.1, that is, the item p levels are not the same, the determination of KR 21 is best exemplified using a different data set.

**Example 5.3:** A 20-item tool is used to assess six students’ scores on a research pretest. Items are scored 1 if correct, and 0 if incorrect. The mean of the resulting test-score distribution is 15 and the variance is 25. KR 21 is determined as follows:

1. There are 20 items on the tool.
2. The mean of the score distribution is 15.
3. The variance of the score distribution is 25.
4. \( KR_{21} = \left( \frac{\text{the number of items}}{\text{the number of items} - 1} \right) \times \left[ 1 - \text{mean of the score distribution} \times \left( \frac{1 - \text{mean of score distribution/number of items}}{\text{variance of the score distribution}} \right) \right] \)

\[
KR_{21} = \left( \frac{20}{19} \right) \left[ 1 - \frac{15}{20} \times \frac{15}{20} \times \frac{25}{20} \right] \\
= (1.05) \left[ 1 - (0.75) \times (0.75) \right] \\
= (1.05) \left[ 1 - (0.5625) \right] \\
= (1.05) (0.4375) \\
= 0.892, \text{ rounded to 0.89}
\]

Additional examples of internal consistency reliability can be found in the studies of Blackwood and Wilson-Barnett (2007); Lavoie-Tremblay et al. (2008); Hatfield (2008); and Rew, Grady, Whittaker, and Bowman (2008).

When a subjectively scored measure is employed, two types of reliability are important—interrater reliability and intrarater reliability.

**Interrater reliability** refers to the consistency of performance (i.e., the degree of agreement) among different raters or judges in assigning scores to the same objects or responses. Thus, interrater reliability is determined when two or more raters judge the performance of one group of subjects at the same time.

To determine interrater reliability, one would:

1. Employ two or more competent raters to score the responses of one group of subjects to a set of subjective test items at the same time.
2. Use an appropriate correlation coefficient to determine the degree of agreement among the different raters in assigning the scores. If only two raters are used, the Pearson product–moment correlation coefficient \( r_{xy} \) may be used as an index of agreement between them. When more than two raters are employed, coefficient alpha may be used, with the column headings representing the judges and the row headings representing the subjects’ performance ratings. Table 5.2 presents an example of alpha employed for the determination of the interrater reliability of six judges of five subjects’ performance on a subjective tool.

An interrater reliability coefficient of 0 indicates complete lack of agreement among judges; a coefficient of 1.00 indicates complete agreement. It should be noted that agreement does not mean that the same scores were assigned by all raters, but rather that the relative ordering or ranking of scores assigned by one judge matches the relative order assigned by the other judges. Interrater reliability is especially important when observational measures are employed as well as when other subjective measures are used, such as free responses requiring categorization, essays, and case studies. Raters are often trained to a high degree of agreement in scoring subjective measures using the interrater reliability procedure to determine when the raters are
using essentially the same criteria for scoring the responses. Examples of interrater reliability can be found in the studies of Arnold (2003); Immers, Schuurmans, and van de Bijl (2005); and Valentine and Cooper (2008).

Intrarater reliability refers to the consistency with which one rater assigns scores to a single set of responses on two occasions. To determine intrarater reliability:

1. A large number of subjects are asked to respond to the same subjective tool.
2. Scores are assigned to the responses using some predefined criteria.
3. Answers are not recorded on the respondents’ answer sheets and anonymity of respondents is protected as much as possible.
4. Approximately 2 weeks after the first scoring, response sheets are shuffled and rescored a second time by the same rater who scored them on occasion one, using the same predefined criteria.
5. The Pearson correlation coefficient ($r_{xy}$) between the two sets of scores is determined as a measure of agreement.

A 0 value for the resulting coefficient is interpreted as inconsistency, and a value of 1.00 is interpreted as complete consistency. Intrarater reliability is useful in determining the extent to which an individual applies the same criteria to rate responses on different occasions and should be employed for this purpose by those who use subjective measures. This technique, because of the time lapse between the first and second ratings, also allows one to determine to some extent the degree to which ratings are influenced by temporal factors. An example of intrarater reliability can be found in Burge (2003).

### TABLE 5.2 Example of Alpha Employed for the Determination of Interrater Reliability for Six Judges’ Rating of Five Subjects’ Performances

<table>
<thead>
<tr>
<th>Subjects</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>10</td>
<td>8</td>
<td>42</td>
</tr>
<tr>
<td>B</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>9</td>
<td>7</td>
<td>35</td>
</tr>
<tr>
<td>C</td>
<td>10</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>8</td>
<td>10</td>
<td>42</td>
</tr>
<tr>
<td>D</td>
<td>10</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>10</td>
<td>8</td>
<td>42</td>
</tr>
<tr>
<td>E</td>
<td>9</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>10</td>
<td>8</td>
<td>42</td>
</tr>
<tr>
<td>Total</td>
<td>47</td>
<td>23</td>
<td>23</td>
<td>21</td>
<td>47</td>
<td>43</td>
<td>204</td>
</tr>
<tr>
<td>Mean</td>
<td>9.4</td>
<td>4.6</td>
<td>4.6</td>
<td>4.2</td>
<td>9.4</td>
<td>8.6</td>
<td>40.8</td>
</tr>
<tr>
<td>Variance</td>
<td>0.64</td>
<td>0.24</td>
<td>0.20</td>
<td>0.70</td>
<td>0.64</td>
<td>1.44</td>
<td>8.56</td>
</tr>
</tbody>
</table>

\[
\alpha = \frac{6}{5} \left[1 - \frac{3.86}{8.56}\right] \\
= (1.2)(1 - 0.45) \\
= (1.2)(0.55) \\
= 0.66
\]
ESTIMATING THE RELIABILITY OF CHANGES IN TEST LENGTH

In many instances, as a result of item analysis that is discussed in Chapter 6, it appears that a measure might be improved either by shortening its length, by eliminating faulty items, or by adding more items. Similarly, when a measure is being developed and tested, the test constructor often will include more items than the number desired for the final form in order to assess by item analysis the performance of individual items with the intent to retain the best items and eliminate faulty items. In these cases, it is important to remember that reliability is a function of test length; that is, a longer test tends to demonstrate a higher reliability than a shorter test. For this reason, following an item analysis, it is often useful to estimate what the reliability of the measure would be if test length were also varied from the form tested. The Spearman–Brown formula permits one to estimate the reliability of a shortened or lengthened measure with known reliability. The assumption, when this approach is used, is that although test length is changed, the nature of the test is not.

Example 5.4: An original 100-item measure has a known reliability of 0.80 and as a result of item analysis, is to be reduced to half its original length or 50 items. The reliability of the shortened version is determined in the following manner. The reliability of the shortened version = the length of the shortened test (1/2) multiplied by the original reliability (0.80) divided by 1 + (the length of the shortened test − 1) multiplied by the original reliability. Specifically,

1. The length of the shortened test is 1/2 multiplied by the original reliability (0.80) = 0.40.
2. 0.40/1 + (1/2 − 1) (0.80) =
3. 0.40/(1/2 − 1) (0.80) =
4. 0.40/1 − (0.40) =
5. 0.40/0.60 =
6. 0.66

The reliability of the shortened version is estimated to be 0.66.

Example 5.5: To estimate the reliability of a measure three times as long as the original measure, with an original reliability of 0.20, the reliability of the lengthened test =

1. Multiple the increased length of the test (3) by the original reliability (0.20) = 0.60
2. Divide 0.60 by 1 + (3 − 1)(0.20) =
3. 0.60/1.40 =
4. 0.42

The reliability of the lengthened test is estimated to be 0.42.

It should be noted that whenever a test is to be lengthened, it is important to consider the potential negative effects of increased test length; that is, extreme increases in test length may introduce unwanted factors, such as boredom, fatigue, diminished
response rate, and other variables, which may actually serve to decrease rather than increase reliability.

**CRITERION-REFERENCED RELIABILITY PROCEDURES**

In criterion-referenced measurement, reliability is concerned with the consistency or dependability with which a measuring device classifies or categorizes phenomena. For this reason, some researchers use the terms *dependability* or *agreement* to refer to reliability of criterion-referenced tests or measures (Brown & Hudson, 2002). In the case of criterion-referenced results, the range of variability is often quite reduced, particularly when scores have been divided into gross categories, such as master and nonmaster. In the norm-referenced case, scores are usually highly variable and reliability is calculated on the basis of parametric correlational analyses. With criterion-referenced measurement the resulting scores are generally less variable than in the norm-referenced case, so reliability is often determined with nonparametric procedures. However, when criterion-referenced scores are reported as percentages, their variability may be similar to those in the norm-referenced case, and most of the procedures used to estimate reliability in the norm-referenced case are also appropriate to assess the reliability of a criterion-referenced measure (Popham, 1978).

In some cases, a criterion-referenced measure may yield scores that are quite variable as far as the actual scores are concerned, but the interpretation of the range of scores would have reduced variability. For example, if a nursing instructor uses a test to determine whether a student has mastered the requisite knowledge in a maternity nursing course, the potential score range might be 0 to 100%. However, assume that the cut score for mastery is set at 80%. If the student scores 75% on the test, the student has not mastered the content. Based on the way in which the scores on the test are interpreted and used, the concern for testing reliability is on the consistency with which the measure classifies the subjects within the specified categories of the content domain. Even if a whole class of 20 students is tested by the measure, with the scores reflecting marked variability, the primary concern would be the consistency with which the measure classifies each student as master or nonmaster in terms of the stated criterion standard, the cut score. This brings to mind another very important point. In the case of criterion-referenced measurement, unless the standard or cut score has high validity, the computation of a reliability index has little significance. A high-reliability index in a situation in which the standard has been improperly set may mean only that the measure consistently classifies objects or phenomena incorrectly.

In the criterion-referenced framework, reliability is usually estimated by employing test–retest, parallel forms, and intrarater and interrater agreement procedures.

**CRITERION-REFERENCED TEST–RETEST PROCEDURE**

The focus of the test–retest procedure for criterion-referenced measures is on the stability over time of the classification of phenomena by a measure on two separate measurement occasions. In other words, the focus is on the ability of a measure to consistently classify objects or persons into the same categories on two
separate occasions. The extent to which a criterion-referenced measure is able to reflect stability of results over time is an indication of the degree to which it is free from random measurement error.

To estimate test–retest reliability for a given criterion-referenced measure, an investigator would follow the same general guidelines in administering the measure as described for the norm-referenced case. However, the calculation of the reliability index would be different because of the difference in the way criterion-referenced test results are interpreted and used.

Two statistics have been identified that may be employed to assess the stability of criterion-referenced test results for the test–retest procedure, regardless of the number of categories established by the measure: $P_o$, also termed percentage agreement or the coefficient of agreement, is the proportion of observed agreements in classifications on both occasions and is the simplest index of agreement (Portney & Watkins, 2009); $K$ or kappa also referred to as Cohen’s $K$, is the proportion of persons consistently classified in the same category on both occasions beyond that expected by chance. Hence, $K$ is $P_o$ corrected for chance agreements.

**Computation of $P_o$ (Percentage Agreement or the Coefficient of Agreement)**

$P_o$ is the number of exact agreements divided by the number of possible agreements. According to Subkoviak (1980), $P_o$ is best computed by the following formula.

**Formula 5.1:** Computation of $P_o$ (percentage agreement; Subkoviak, 1980)

$$P_o = \frac{1}{m} \sum_{k=1}^{m} P_{kk}$$

where $m$ = the number of classification categories, and $P_{kk}$ = the proportion of objects or persons consistently classified in the $k$th category.

For illustrative purposes, assume that a criterion-referenced measure designed to assess a nurse’s attitudes toward elderly clients is administered to 30 nurses at 2-week intervals to determine test–retest reliability. Results are illustrated in Table 5.3. The $P_o$ would be the proportion of student nurses consistently classified

<table>
<thead>
<tr>
<th>TABLE 5.3 Hypothetical Test Results Matrix for 30 Nurses for Computing $P_o$ and $K$ on a Measure of Nurse Attitudes Toward Elderly Clients</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>First Administration</strong></td>
</tr>
<tr>
<td><strong>Positive</strong></td>
</tr>
<tr>
<td>Positive</td>
</tr>
<tr>
<td>Negative</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
</tr>
<tr>
<td><strong>Second Administration</strong></td>
</tr>
<tr>
<td><strong>Positive</strong></td>
</tr>
<tr>
<td>Negative</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
</tr>
</tbody>
</table>

©Springer Publishing Company
with positive/positive and negative/negative attitudes on both testing occasions. Thus, \( P_o \) would be the total proportion of the values in cells A and D. Hence:

\[
P_o = \frac{15}{30} + \frac{12}{30} = 0.50 + 0.40 = 0.90
\]

Therefore, in this example 0.90 or 90% of the classifications made by the measure on both testing occasions were in agreement. However, some small portion of this estimate can be attributed to chance and 0.90 is, therefore, an overestimate of the stability of the test.

The proportion of chance agreements (\( P_c \)) for the data in Table 5.3 can be computed by the product of the corresponding row and column totals as indicated by Formula 5.2.

**Computation of the Proportion of Chance Agreements**

Chance agreement is reflected by the number of expected agreements divided by the number of possible agreements. Subkoviak (1980) provides the following formula for calculating chance agreements:

**Formula 5.2:** Calculating one proportion of chance agreements (\( P_c \)) (Subkoviak, 1980)

\[
P_c = \sum_{k=1}^{m} P_k P_k
\]

where \( m \) = the number of classification categories, and \( P_k P_k \) = the proportion of objects or persons assigned to category \( k \) on each measurement occasion, respectively.

In this situation \( P_c \) would be computed by the proportions for \((A + B)(A + C) + (C + D)(B + D)\). Thus,

\[
P = \frac{(17)(16)}{30} + \frac{(13)(14)}{30} + \frac{(13)(14)}{30} + \frac{(17)(16)}{30}
\]

\[
= (0.57 \times 0.53) + (0.43 \times 0.47)
\]

\[
= 0.30 + 0.20
\]

\[
= 0.50
\]

**Computation of Cohen’s Kappa (\( K \))**

The proportion of nonchance agreements is provided by kappa (\( K \)) (Cohen, 1960). \( P_o \), observed agreements, and \( P_c \), chance agreements, are used to calculate \( K \) as follows:

**Formula 5.3:** Calculating the proportion of nonchance agreements (\( K \)) (Martuza, 1977; Subkoviak, 1980)

\[
K = P_o - P_c
\]
In the present example, $K$ is computed by:

$$K = \frac{0.90 - 0.50}{1 - 0.50} = \frac{0.40}{0.50} = 0.80$$

**Computation of $K_{max}$**

An upper-bound $K$ value of 1.00 will result only when the marginal distributions for the two administrations have the same shape or proportions in them, for example, when the proportions in the right upper cell ($A + B$) and in the bottom left cell ($A + C$) of the table are the same. One can determine the maximum possible value of $K$ for a specific situation by adjusting the values within the cells of the table (cells $A$, $B$, $C$, and $D$) to reflect the maximum number of possible agreements or consistent test classifications that could be congruent with the observed marginal proportions (marginal proportions are not changed) and by calculating a revised version of $K$ using the adjusted values. When this is done the resulting value is $K_{max}$, which represents the upper limit value that $K$ could take on with the particular distribution of results. The $K_{max}$ value provides information that can facilitate a better interpretation of a specific $K$ value. When the $K/K_{max}$ ratio is calculated, it provides a value that can be interpreted on a standard scale. The upper limit of this ratio is 1.00. The closer this ratio is to 1.00, the higher the degree of consistency of classifications from administration 1 to administration 2.

The computation of $K_{max}$ for the information provided in Table 5.3 is conducted by first transforming all of the values in Table 5.3 to proportions as shown in Table 5.4. The proportions in the cells of Table 5.4 are then adjusted to reflect the maximum number of agreements in cells $A$ and $D$ that could possibly be congruent with the observed marginal proportions as shown in Table 5.5. At this point $K_{max}$ can be calculated using the formula for $K$. Hence:

$$K_{max} = \frac{-0.53 + 0.43 - [(0.57)(0.53) + (0.43)(0.47)]}{1 - (0.57)(0.53) + (0.43)(0.47)}$$

$$= \frac{0.96 - 0.50}{1 - 0.50} = \frac{0.46}{0.50} = 0.92$$

The $K$ value for the present example is 0.80 and the $K_{max}$ value is 0.92. Based on these findings it can be assumed that the measure classified the nursing students with a relatively high degree of consistency, since $K/K_{max}$ is 0.87.
TABLE 5.4 Data from Table 5.3 Expressed as Proportions

<table>
<thead>
<tr>
<th></th>
<th>Positive</th>
<th>Negative</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>First Administration</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>(A) 0.50</td>
<td>(B) 0.07</td>
<td>(A + B) 0.57</td>
</tr>
<tr>
<td>Negative</td>
<td>(C) 0.03</td>
<td>(D) 0.40</td>
<td>(C + D) 0.43</td>
</tr>
<tr>
<td><strong>Second Administration</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Totals</td>
<td>(A + C) 0.53</td>
<td>(B + D) 0.47</td>
<td>(A + B + C + D) 1.00</td>
</tr>
</tbody>
</table>

TABLE 5.5 Adjustments Required in Table 5.4 for the Calculation of $K_{\text{max}}$

<table>
<thead>
<tr>
<th></th>
<th>Positive</th>
<th>Negative</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>First Administration</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>(A) 0.53</td>
<td>(B) 0.04</td>
<td>(A + B) 0.57</td>
</tr>
<tr>
<td>Negative</td>
<td>(C) 0.00</td>
<td>(D) 0.43</td>
<td>(C + D) 0.43</td>
</tr>
<tr>
<td><strong>Second Administration</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Totals</td>
<td>(A + C) 0.53</td>
<td>(B + D) 0.47</td>
<td>(A + B + C + D) 1.00</td>
</tr>
</tbody>
</table>

Criterion-Referenced Parallel-Forms Procedure

There are some situations when more than one form of a measure is desirable. For instance, in situations in which subjects are measured before and after a nursing intervention, it may be preferable to administer a parallel form on subsequent administrations. The test–retest procedure has a potential pitfall, which makes that approach to the study of the reliability questionable, as significant events may occur during the between-testing interval that might interfere with results on the second administration. Also, administration of the test on the first occasion might influence the results on the second testing, particularly if the same measure is used. The use of parallel forms of a measure could help remedy such situations. However, a major concern in instances in which parallel forms of a measure are used is whether the two
forms produce a substantial degree of agreement or consistency in the classification of subjects in a specified group.

Two criterion-referenced measures are considered parallel if they assess the same content domain, that is, if they were constructed with the same set of test specifications and if items are relatively homogeneous. Parallel forms of a criterion-referenced measure may be created through random item selection from a pool of items constructed with the same test specifications or the same item-generation rules (Popham, 1978).

The approach to the estimation of the reliability of parallel forms involves administering the two alternate forms of the measure to one specific group on the same measurement occasion. If it is not possible to administer the two forms of the test at the same time, they should be administered under similar circumstances within a short period of time. After the two versions of the measure are administered, $P_o$ and $K$ would be calculated in the same manner used in the test–retest case. Data from the two forms would be compiled and placed in a matrix as shown in Table 5.3. However, the label “First Administration” should be changed to “Form 1,” and “Second Administration” to “Form 2.” If the two forms have high parallel-form reliability, there will be a high consistency in the classification of subjects into categories. In the parallel-forms procedure, high $P_o$ and $K$ values reflect consistency between the alternate forms of a measure.

**Criterion-Referenced Interrater and Intrarater Agreement Procedures**

As with the test–retest and the parallel-forms procedures, $P_o$ and $K$ can be employed to estimate interrater and intrarater agreement, which also may be referred to as interjudge and intrajudge agreement (Tindal & Marston, 1990). The focus of interrater agreement in the criterion-referenced case is on the consistency of classifications of two (or more) different raters who classify a specified group of objects or persons using the same measurement tool on the same measurement occasion. For example, if a rating tool designed to measure the environmental safety of nursing units is used, two nurse raters could be employed to independently classify a group of nursing units as either safe or unsafe one at a time. Once results are obtained, $P_o$ and $K$ could be calculated to determine interrater agreement for the classification of the safety of the nursing units. The values of $P_o$ and $K$ are computed in the same manner as indicated previously and used as the index of interrater agreement. Prior to computing $P_o$ and $K$ the data would be set up in a matrix table similar to Table 5.3, but with the appropriate label changes, that is, changing “First Administration” to “Rater 1,” “Second Administration” to “Rater 2,” “Positive” to “Safe,” and “Negative” to “Unsafe.”

Intrarater agreement for criterion-referenced measurement situations is the consistency with which a single rater classifies a group of persons or objects, using a specified measuring tool after rating each person or object on two separate occasions. In instances when intrarater agreement is used, there is a danger that the two separate rating situations are not consistent with each other unless the situation has been captured by a recording device, such as video or audio recordings or written documents from which both ratings can be made. Another danger is that the first rating might affect the second rating. Steps would be taken to minimize this problem by using such techniques as obscuring the identification of the persons or objects being rated,
altering the order in which ratings are done, and reordering the pages of the rating tool if appropriate. Data are arrayed in a matrix in the manner discussed previously, with the proper labeling changes. $P_o$ and $K$ are then calculated to provide an index of intrarater agreement (Martuza, 1977).

In cases where criterion-referenced measurement is based on ratings by observers, there are several types of rating errors that can negatively impact the reliability and, thus the validity of ratings. These include error of standards, halo error, logic error, similarity error, and central tendency error (Shrock & Coscarelli, 2007).

*Error of standards* result when numerical and descriptive rating scales fail to provide definitions of behaviors specific enough to prevent raters from using their own standards to rate items differently from those intended by the developer of the measure. Hence, different raters would be more likely to rate the same items differently, thereby reducing reliability of ratings among raters.

*Halo error* occurs when raters allow their opinion of the performer to influence their performance ratings. This may be done subconsciously and may be in a negative or positive direction. Therefore, halo error can affect scores negatively or positively.

*Logic error* results when a rater rates one characteristic of a performer when another characteristic is supposed to be the focus of the rating. This occurs when the rater is not fully aware of the independence of two performance characteristics. For example, suppose a clinical instructor assumes that a student who takes longer to perform a procedure is less knowledgeable about the procedure. In actuality, speed of conducting the procedure may have nothing to do with the amount of knowledge the student has about the procedure.

*Similarity error* occurs when raters have the tendency to rate a performer that they perceived to be similar to them more highly than those who they perceive are “different.” This error also is referred to as “similarity-to-me error.”

*Central tendency error* is associated with rating scales that allow raters to choose points along a continuum, such as with behaviorally anchored, descriptive, or numerical scales. Raters will often avoid rating performers on the extreme anchors of rating scales and they tend to group ratings more in the middle of the scale. This behavior is so consistent that for most Likert-type scales the extreme positions are often lost; thus, a seven-point scale is responded to as if it is a five-point scale during ratings. For this reason, some psychometricians recommend using only an even number of categories on rating scales (Shrock & Coscarelli, 2007).

**Weighted Kappa ($K_w$)**

When Cohen’s kappa is calculated, the frequencies along the agreement diagonal are used. This approach assumes that all disagreements are of equal seriousness although this could not in actuality be the case. In some instances, the researcher may want to assign greater weights to some disagreements in relation to others due to the differential risks of some disagreements. For example, when criterion-referenced scores are assigned using ordinal categories with more than two categories, such as 0, +1, +2, or +3, a disagreement between ratings of 0 and +3 is much more serious than a disagreement between 0 and +1. In the former situation there is a difference of three categories
of disagreement and in the later a difference of only one category. Therefore, the disagreement in the former case is much more serious than the later. In situations where disagreements can be differentiated in this way, a weighted kappa, $K_w$, can be used to estimate reliability (Cohen, 1968; Portney & Watkins, 2009). A weighted kappa gives more credits to the more serious disagreements than those that are less serious by assigning different weights to the off-diagonal cells based on the seriousness of disagreements. The assignment of weights is basically a judgmental process (Cohen, 1968), but should be consistent with a hypothesis that defines the relative seriousness of the misclassifications or disagreements. The assignment of weights can be incremental, asymmetrical, or symmetrical weights. 

Incremental weights assume that the categories are on an ordinal scale with equal weights; asymmetrical weights are based on the assumption that disagreements do not fit a uniform pattern but that the direction of disagreement is important; and, symmetrical weights assume that the direction of disagreement is unimportant.

Weights for incremental disagreements can be assigned by using the formula: $w = (r_1 - r_2)^2$; where $w$ = the assigned weight and $r_1$ and $r_2$ = scores assigned by rater 1 and rater 2 to a cell (Fleiss & Cohen, 1973). Hence, $r_1 - r_2$ indicates the deviation from agreement for each cell in the agreement matrix. Weights of zero would automatically be assigned to all cells on the diagonal where there is no disagreement.

When assigning asymmetrical weights the evaluation of weights do not fit a uniform pattern. It is important to consider if the direction of the disagreement is important for assigning weights for disagreements and the amount of weight. For example, let’s assume two raters have the task of rating the developmental behaviors of 2-year-old boys as either “below normal,” “normal,” or “advanced.” Assigning a rating of “advanced” to a child who is “below normal” would be considered a more serious disagreement than assigning a rating of “normal” to a child who is “advanced.” This would be the case because needed corrective intervention may not be given to the child who is “below normal” who has been misclassified as “advanced.” It also would be of concern if there was a misclassification of a child who was “below normal” as “normal.” On the other hand, the misclassification of a child who was “advanced” as “normal” would not be quite as serious. Therefore, the more serious the hypothesized misclassification, the more heavily the misclassification would be weighted.

In the case of assignment of symmetrical weights, the direction of disagreement is considered unimportant. For example, one might argue that any disagreement between “below normal” and “advanced” is twice as serious as a disagreement between “below normal” and “normal,” and that a disagreement between “normal” and “advanced” is only minimally important. Therefore, disagreements between “below normal” and “advanced” might be assigned as weight of 6, whereas a disagreement between “below normal” and “normal” would be given a score of 3, and a weight of 1 might be assigned to disagreements between “normal” and “advanced.”

Calculation of $K_w$ would proceed with the use of the assigned weights to each cell in the agreement matrix. A limitation of this procedure is that if weights are assigned in an arbitrary manner and are not based on a sound hypothesis, then the value of $K_w$ would be arbitrary also. Therefore, the rationale for the assignment of weights
must be clear and scientifically defensible. Hence, when \( K_w \) is calculated the research report should specify the weighting procedure and the rationale for the assignment of weights.

**Interpretation of \( P_o \) and \( K \) Values**

Several points should be kept in mind regarding the interpretation of \( P_o \) and \( K \) values. The value of \( P_o \) can range from 0 to 1.00. Total disagreement in observed test classifications is reflected by a \( P_o \) value of 0, whereas total agreement in observed results is reflected by a \( P_o \) value of 1.00. As indicated by the formula for \( K \), the value of \( K \) is always less than or equal to \( P_o \). The size of the difference between \( P_o \) and \( K \) is always a function of the size of \( P_o \) or chance agreements. The value of \( K \) always lies within an interval between −1.00 (which represents complete inconsistency of test results) and 1.00 (which reflects total consistency of results; Hashway, 1998). The upper limit of 1.00 for \( K \) is fixed; however, the lower-bound value may fluctuate from one situation to another depending upon several influencing factors. Landis and Koch (1977) proposed the standards for interpretation of the strength of agreement for \( K \) as follows: Kappa values above 0.80 indicate near perfect reliability, values between 0.61 and 0.80 imply substantial reliability, values between 0.41 and 0.60 indicate moderate reliability, kappas between 0.21 and 0.40 reflect fair reliability, and values equal to or less than 0.21 indicate slight reliability. Both \( P_o \) and \( K \) are affected by factors such as test length, number of response alternatives (e.g., when items are multiple choice), the value of the cut score used to classify subjects, sample size, and the homogeneity of the group of subjects. For example, as the number of response categories increases, the extent of agreement will usually decrease, and homogeneous samples often lead to a high percentage of agreements. Small samples can result in misleading results due to rater bias (Portney & Watkins, 2009). At this time, guidelines related to these factors (which would facilitate further interpretation of \( P_o \) and \( K \) values) have not been explained. Whenever \( P_o \) and \( K \) are used to describe the reliability of a criterion-referenced test, these influencing factors should be clearly described because of their impact on the values of \( P_o \) and \( K \). However, it should be noted that the size of the difference between \( P_o \) and \( K \) represents the amount of susceptibility of the decision process to chance factors (Martuza, 1977). It is possible to use kappa with more than two raters or more than two test–retest administrations (Fleiss, 1971); however, it is best to use separate kappas computed for pairs of raters or pairs of test–retest administrations to facilitate ease of interpretation of results.

**SUMMARY**

Every measurement involves some error that cannot be eliminated but can be reduced by using sound approaches to measurement. Random errors of measurement affect reliability. Reliability must be assessed every time a given measure is employed. The domain-sampling model is the conceptual basis of choice for the determination of reliability in the norm-referenced case. Norm-referenced reliability is usually estimated using a test–retest, parallel form, and/or internal consistency procedure. In addition, when a subjectively scored measure is used, it is important to consider...
interrater and/or intrarater reliability as well. When variations in the length of a measure result from item analysis, estimations of reliability using the Spearman-Brown formula should be considered prior to actually making such modifications.

Each of the types of reliability procedures is applicable in both norm-referenced and criterion referenced measurement. The principles for each type of reliability assessment are the same regardless of the measurement framework employed. However, the approach to calculation of the reliability coefficient depends upon the nature of the score that results. As noted previously, because norm-referenced measures result in scores that are at the interval level of measurement, parametric statistical procedures are used. In the criterion-referenced case, nonparametric statistics are employed when categorical scores or classifications result and parametric statistics are permissible when percentage scores result and the distribution of scores are not highly skewed. Table 5.6 summarizes the types of reliability procedures and related statistics that can be applied in the norm- and criterion-referenced cases.

When the criterion-referenced measurement framework is used, reliability procedures follow the same general principles as in the norm-referenced case. However, scores are often not as variable because, in many cases, criterion-referenced measures result in classifications or the placement of the objects that are the focus of measurement into categories. In this instance, percentage agreement ($P_o$) and/or kappa ($K$) are typically used as approaches to the calculation of the various types of reliability. When percentage scores are used with criterion-referenced measures, it may be appropriate to use the same statistical approaches as in the norm-referenced case if scores are not highly skewed.

| TABLE 5.6 Types of Norm-Referenced and Criterion-Referenced Reliability-Procedures-Related Statistics |
|---------------------------------------------------------------|---------------------------------------------------------------|
| Reliability Procedure                                      | Norm-Referenced Statistic(s)                                 | Criterion-Referenced Statistic(s)                           |
| Test–retest: Concerned with consistency of measurements that one instrument or tool elicits from one group of subjects on two separate measurement occasions (stability assessment). | Correlation of the two sets of scores using Pearson product–moment correlation. | The percentage agreement ($P_o$) of classification of subjects on the two separate measurement occasions is calculated. Kappa ($K$), which is ($P_o$) adjusted for chance agreements, also can be computed. Pearson product–moment correlation can be used with percentage scores that are not highly skewed. Nonparametric correlation procedure, that is, Spearman rho, can be computed if data are highly skewed. |

(continued)
### TABLE 5.6 Types of Norm-Referenced and Criterion-Referenced Reliability-Procedures-Related Statistics (continued)

<table>
<thead>
<tr>
<th>Reliability Procedure</th>
<th>Norm-Referenced Statistic(s)</th>
<th>Criterion-Referenced Statistic(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parallel forms procedure: Concerned with assessing the consistency of measurements that alternate forms of an instrument or tool elicit from the same group of subjects during a single administration (equivalence assessment).</td>
<td>Correlation of two sets of scores obtained from each tool using Pearson product–moment correlation. Two norm-referenced measures are considered parallel if: (a) they were constructed using the same objectives and procedures; (b) have approximately equal means; (c) have equal standard deviations; and (d) have equal correlations with any third variable.</td>
<td>The percentage agreement ( (P_o) ) in the classification of subjects by the two parallel measures is computed. Kappa ( (K) ) also can be computed. Two criterion-referenced measures are considered parallel if: (a) they were constructed with the same set of test specifications; and (b) items are relatively homogeneous in nature. Pearson product–moment correlation or Spearman’s rho can be used with percentage scores as appropriate.</td>
</tr>
<tr>
<td>Interrater reliability: Concerned with consistency of performance (i.e., degree of agreement) among different rater or judges in assigning scores to the same objects or behaviors in the same tool and/or the same predefined criteria (equivalence assessment).</td>
<td>Correlation of the two sets of scores obtained from each rater or judge using Pearson product–moment correlation.</td>
<td>The percentage agreement ( (P_o) ) of the classification of subjects by the two raters or judges is computed. Kappa ( (K) ) also can be computed. Pearson product–moment correlation or Spearman’s rho can be used with percentage scores as appropriate.</td>
</tr>
<tr>
<td>Intrarater reliability: Refers to the consistency with which one rater assigns scores to a set of behaviors on two occasions (observed under the same conditions) using the same instrument and/or the same predefined criteria (stability assessment).</td>
<td>Correlation of the two sets of scores obtained from one rater or judge for the two occasions using Pearson product–moment correlation.</td>
<td>The percentage agreement ( (P_o) ) of the two sets of classifications obtained from a single rater or judge is computed. Kappa ( (K) ) also can be calculated. Pearson product–moment correlation or Spearman’s rho can be used with percentage scores as appropriate.</td>
</tr>
</tbody>
</table>

(continued)
### TABLE 5.6 Types of Norm-Referenced and Criterion-Referenced Reliability-Procedures-Related Statistics (continued)

<table>
<thead>
<tr>
<th>Reliability Procedure</th>
<th>Norm-Referenced Statistic(s)</th>
<th>Criterion-Referenced Statistic(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal consistency reliability: Concerned with the consistency of performance of one group of individuals across the items on a single instrument. Alpha is equal in value to the mean of the distribution of all possible split-half coefficients associated with a specific set of test or questionnaire data (internal consistency assessment).</td>
<td>The alpha coefficient is calculated. KR 20 and KR 21 are special cases of alpha, which are used when data are dichotomously scored. KR 20 is used when the item difficulty levels cannot be assumed to be the same. KR 21 can be used when item difficulty levels are assumed to be the same.</td>
<td>KR 20 and KR 21 may be calculated for some criterion-referenced measures when data are not highly skewed and are dichotomously scored, e.g., when each item is scored 1 if correct and 0 if incorrect or missing.</td>
</tr>
</tbody>
</table>

**Note:** The closer the correlation coefficient, alpha, $\rho$, or $K$ are to 1.00 (or 100%), the more reliable the measure or tool is presumed to be.

### REFERENCES


206  II  ■  Understanding Measurement Design


Qualitative inquiry is an approach to research that focuses on the description and understanding of phenomena (or concepts) within the social world from the perspective of individuals who are experiencing that world. Nursing's unique disciplinary focus on interactions among dimensions of the person, health, environment, and the process of nursing positions nurses to explore and disseminate a deeper understanding of the illness experiences in order to promote higher levels of wellness. Thus, qualitative approaches are very appropriate for nurse scientists because this class of methods enables a deeper understanding of phenomena of interest, such as the processes associated with health and illness (e.g., grounded theory); how culture influences people's lives, health patterns, and responses to illness (e.g., ethnography); and the essential meaning ascribed to living with a certain state of health or illness experience (e.g., phenomenology). The knowledge developed through rigorous qualitative inquiries is integral to advancing nursing science toward the goal of improving health outcomes by supporting others through similar situations (see Morse, 2008a).

Over the past 40 years, as qualitative methods have become an accepted means of studying phenomena of interest to nursing, greater attention has been paid to methodological integrity and rigor. This has occurred for a number of reasons, but perhaps most important, the advancement of nursing science has become more focused. Moving nursing science forward requires more of researchers than simply filling our journals with interesting yet merely descriptive qualitative studies. Today's qualitative researchers must be cognizant of how their project contributes to theory development and the advancement of nursing science toward promoting the health and well-being of the persons for whom we care (Morse, 2012).

Stop to consider the emphasis on evidence-based or research-driven practice and translational research. Couple these initiatives with tighter research budgets, and the need for strategically executing qualitative research with the goal of advancing the science (and, therefore, practice) of nursing becomes apparent (see Sandelowski & Leeman, 2012, for a more detailed discussion of the utility of qualitative research). Although there is certainly a place in nursing science for high-quality descriptive work, this is a limited niche. It is imperative for nurse researchers to begin a project with a thorough understanding of the state of the science so that subsequent projects may
be designed to specifically address identified gaps in understanding that preclude meeting the goal of facilitating higher levels of wellness through nursing interventions. Even in this climate of budgetary constraints and emphasis on outcomes, high-quality qualitative studies continue to merit highly competitive grant awards (see Federal RePORTER [Research Portfolio Online Reporting Tools]).

From an ontological and epistemological perspective, this evolution may be pushing qualitative research, as a whole, toward a more positivistic paradigmatic perspective (see Angen, 2000). However, in order for qualitative research to be funded and, postfunding, for the findings to be useful to nursing science, issues, such as validity, reliability, and generalizability of the research, must be addressed (Morse, 1999a). There are excellent nurse researchers who hold opposing paradigmatic stances that are incongruent with these notions (e.g., an autoethnographer who espouses contextually bound cocreated truth); however, major funding sources typically require that all researchers address these issues regardless of whether a qualitative or quantitative design is proposed. (For further information, see Sandelowski & Barroso [2003] guidelines for qualitative research proposals.) This chapter presents common threats to validity, reliability, generalizability, and ways to minimize these threats in qualitative studies while introducing other terms that may be used by the qualitative researcher (e.g., trustworthiness instead of validity).

OVERVIEW OF QUALITATIVE RESEARCH

Qualitative research is undertaken for a number of reasons. Most often, qualitative methods are applied when little is known about a concept or phenomenon as a result of limited previous research (Morse, 2012). This scenario is common when the topic is a new area of interest in the science, or when prior research and/or theory do not fit the phenomenon of interest because of an evolution in the context of the experience (e.g., changed care delivery systems, new treatments, changed societal views, or expanded cultural views). In these cases, qualitative research could be done to describe the phenomenon more fully or to develop or refine extant theory to better capture relevant experiences of those people living through the phenomenon. There are numerous characteristics of qualitative research that distinguish it from quantitative research. The following is a list of some of the hallmarks of a qualitative approach:

- The dominant mode of reasoning in qualitative research is inductive: The researcher examines specific instances (i.e., data) in order to derive more general conclusions (i.e., findings). In other words, qualitative researchers interpret designated units of data in order to build descriptive characteristics, categories, concepts, or models from these data. Although confirming and refuting conjectures (or hypotheses) is typically a part of most analytic strategies, this form of deductive reasoning is not dominant. This hallmark demands cautionary integration of how a selected theoretical framework informs a project; inductive reasoning must be apparent.
- Qualitative research is concerned with process and meaning from an emic (participant’s) perspective. The explicit purpose of most qualitative studies is to explore how individuals make sense of their world and how this ascribed
meaning influences their health, not to show causality. This notion requires the researcher to interpret multiple views of reality into some semblance of pattern (or theory that holds utility for nurses).

- Qualitative perspectives are sensitive to context; therefore, these studies are typically undertaken in the field or naturalistic setting. There is no effort to control or manipulate central or extraneous variables because these variations are intertwined with the phenomenon under investigation. Complex interactions are important to understanding the process and meaning inherent to a phenomenon of interest. Well-designed studies include sampling strategies that maximize the phenomenon of interest and address identified variations through sequential sampling strategies or a clearly articulated delimitation of findings. This hallmark is directly related to the scope of the project: Small projects typically cannot fully explore all identified variations, yet it is critical that unexplored variations not be ignored. In order to advance science, these limitations must be disseminated to permit further strategic investigation of the phenomenon. In qualitative approaches, the researcher is a tool, guiding data collection and analysis. This requires the researcher to continually differentiate between his or her own reality and that of the participant and still be close enough to the participant’s world to understand and describe it (Creswell, 2013; Hegelund, 2005; Morse & Field, 1995; Polit & Beck, 2012). The concept of bracketing is often used to describe a process of setting aside one’s preconceptions that may misdirect the evolving qualitative inquiry; however, the notion of bracketing is often diluted (Gearing, 2004). The use of audit trails may be useful in demonstrating key decisions that guide the project and may be useful in exposing how preconceptions may influence the course of a project (Houghton, Casey, Shaw, & Murphy, 2013).

Although there are a number of qualitative approaches used by nurses, the most commonly used methods are used as examples in this chapter. These include phenomenology, grounded theory, ethnography, focus group, case study, and narrative/storytelling. Each is discussed briefly in this chapter.

**PHENOMENOLOGY**

Phenomenology is a qualitative method that is based in philosophy. Its aim is to understand and then describe in detail the essence of the experience as it is lived by the person (Creswell, 2013; Munhall, 2012; Van Manen, 1990). The goal is to accurately describe the phenomenon under study and to transform the lived experience into a textual expression of its essence (this can be done with text, pictures, music, etc.). The phenomenon is described as it is experienced by the person (e.g., the experience of waiting for the wife of a critically ill patient), without the guidance of theories and, to the extent possible, without researcher preconceptions and presuppositions. Several sources of data are used, including multiple in-depth conversations with the participant(s), so that the role of researcher in data collection is instrumental. Integration of what is known (i.e., theory) is often woven into the final text to further explicate the meaning of an experience.
GROUND THEOR Y

The methods of grounded theory come from a sociological perspective. They were initially develop by Glaser and Strauss (1967) and subsequently refined by Glaser (1978) and Strauss and Corbin (1998). The primary goal of a grounded theory is to generate explanatory models of human behavior that are grounded in the data. Although most grounded theory is based on a thorough understanding of what is known about a phenomenon, preexisting theory is not used to delimit the focus of the inquiry so that the researcher remains open-minded about what concepts will emerge and how they will be organized. A grounded theory documents processes (e.g., the development of trust in health care providers) and change over time to link categories and to develop models. According to Glaser (1978, 1992), the resultant theory must fit the data, work, have relevance, and be modifiable.

Data are collected using interviews, observations, and field notes. A constant comparative process is used for data collection and analysis. Here data are collected and analyzed simultaneously. Thus, the researcher is observing, collecting data, organizing and analyzing the data, and forming theory from the data all at the same time. Hypotheses are compared and tested with incoming data (every piece of information is compared to every other). Within this process, the researcher also uses theoretical sampling, in which the researcher decides what data to collect next and where to find it based on the needs of the developing theory.

ETHNOGRAPHY

Ethnography is a branch anthropology that focuses on the study of cultural groups. The purpose of ethnography is to tell the story of participants’ lives from the perspective of the culture of which they are a part (Fetterman, 1989). This investigation involves an in-depth study of the members of the culture through participant observation, interviews, and field notes (Wolf, 2007). The researcher needs to spend time (or may live) with the group and become part of the cultural setting in order to collect data and understand the cultural group under investigation.

A focused ethnography is a variation of the traditional ethnography and is an approach that may be more appropriate to nursing. Here the participants are linked by location (e.g., a critical care unit) not a place of residence or culture in the anthropological sense, but share behavioral norms and a common language (Morse & Field, 1995; Polit & Beck, 2012). In a focused ethnography, the topic is selected prior to data collection, whereas in a traditional ethnography it emerges from the data. Data collection, including interviews and observations, is limited to particular events and topics related to the event. The product is an understanding of essential culture schemas in relation to the phenomenon of interest (e.g., dying in a critical care unit).

FOCUS GROUPS

A focus group is a technique for data collection that uses group interactions to obtain an understanding of participants’ experiences and beliefs (Krueger & Casey, 2015). A focus group can be used solely as technique to collect data (e.g., within an ethnographic study or grounded theory) or as a stand-alone method. Because in-depth data
cannot be obtained from each participant, focus groups typically are used for problem identification, program planning, and program evaluation. The group is researcher controlled in that the topic chosen and questions asked are those of the researcher. However, the discussion and group diversity and consensus come from the group and its discussion (Morgan, 1997).

**CASE STUDY**

Case studies are an exploration of a “bounded system” or a case (or multiple cases) over time through detailed, in-depth data collection involving multiple sources of information such as interviews, observations documents, archival records, artifacts, and audiovisual materials (Yin, 2014) that are rich in context (Creswell, 2013). The case can be an individual, event, program, or organization. Case studies may also be described in terms of the intent of the analysis. Three variations exist: the single instrumental case study, the collective (or multiple) case study, and the intrinsic case study (Stake, 1995). The product of a case study is a detailed description of the case, including chronology, naturalistic generalizations, and lessons learned from the case (Creswell, 2013).

**NARRATIVE/STORYTELLING**

A narrative is a method that is used “to give a voice to the silenced” (Frank, 1995) and as a way to study humans’ experiences within the social world. Within narrative inquiry, information is gathered for the purpose of storytelling. It is an art of both listening to the narrative and then telling the story (or writing a narrative of the experience; Riessman, 2008). Data are collected through in-depth interview, field notes, journals, letters, and stories told by the participant (Duffy, 2012). The researcher then writes the narrative of the participant (this may be a collaborative effort between the researcher and participant).

**VALIDITY, RELIABILITY, AND GENERALIZABILITY OF QUALITATIVE FINDINGS**

As suggested earlier, in order for qualitative inquiry to advance nursing science, methodological and analytic issues, such as validity, reliability, and generalizability, need to be addressed when a study is designed (see Silverman, 2013). Threats to these issues can occur when the study is initially designed (e.g., picking the wrong method for the research question, or not adequately planning the sample), during data collection and during data analysis. In this section, these issues are defined. This is followed by a discussion of the threats to validity and reliability during sampling, data collection, and data analysis and how these threats can be minimized.

**VALIDITY**

Issues surrounding validity in qualitative research continue to evoke dialogue and disagreement among methodologists (see Sparkes, 2001), particularly from epistemological, ontological, and broader paradigmatic perspectives (Angen, 2000;
Cho & Trent, 2006; Koro-Ljungberg, 2008; Pyett, 2003). For purposes of this discussion, validity is defined as the “truth value,” or trustworthiness of the data and resultant analysis and interpretation, or the extent to which the findings represent reality (Morse & Field, 1995). Miles, Huberman, and Saldaña (2014) extend the notion to address internal and external validity; however, Maxwell (1992) asserts that internal and external validity reflects positivistic assumptions and posits six types of validity relevant to qualitative inquiry. Each of these types of validity (internal validity, external validity, descriptive validity, interpretive validity, theoretical validity, and evaluative validity) is discussed in the following (generalizability is discussed separately).

**Internal Validity**

Internal validity asks the question of whether the researchers are measuring or observing what they think they are measuring or observing (Schensul, Schensul, & LeCompte, 1999). In qualitative terms, Miles et al. (2014) also refer to internal validity as credible or authentic, in which the findings need to make sense and be credible to both the participants and readers. Internal validity is enhanced during both data collection and analysis when the researcher uses multiple sources of data, links data to the emerging categories and theory, confirms findings with additional data, and looks for negative evidence (tries to find data that may not support the analysis/hypotheses being developed).

**External Validity**

External validity asks the question whether the findings and conclusions of the study can be transferred to other contexts, thus generalizable beyond the present study (applicable across groups; Schensul et al., 1999; Miles et al., 2014). Other terms used for external validity are transferability and fittingness (Miles et al., 2014). External validity is enhanced by having an adequate sample, including sample size, sample diversity, and appropriate purposive/theoretical sampling. The findings should be abstract enough to apply to other contexts and contain enough rich description for the reader (or other researchers) to evaluate the findings.

**Descriptive Validity**

Descriptive validity is related to the “truth value,” or credibility/authenticity (or valid description) of the data and what was reported from the data (Maxwell, 1992). Here, the researcher needs to carefully collect and corroborate data (this can be done by obtaining feedback about the accuracy of the data and analysis from the original participants or by using secondary participants to confirm the emerging analysis), and present an accurate account of these data.

**Interpretive Validity**

Interpretive validity relates to meaning or the interpretation of the data (Maxwell, 1992). Does the researcher accurately understand and portray the participant's (emic) view or meaning?
Theoretical Validity

Theoretical validity moves beyond the descriptive and interpretative types of validity to analysis of the validity of a theory (Maxwell, 1992). Thus, does the derived model fit with the data and is it abstract enough to extend the theory beyond description?

Evaluative Validity

Evaluative validity applies an evaluative framework to the objects of the study, not the study design itself (Maxwell, 1992). Here, a judgment is made about the correctness or worth of the meanings or actions. Therefore, how the researcher describes, interprets, or constructs the story or theory is important.

RELIABILITY

Reliability in qualitative research is concerned with consistency over time and across researchers and settings, and objectivity and confirmability (Miles et al., 2014). There is a distinction made between internal and external reliability. Internal reliability is defined by Schensul et al. (1999) as “the degree to which other researchers, given a set of previously generated constructs, would match them with the data in the same way as did the original researcher” (p. 275). Other terms for internal reliability are dependability and auditability. Miles et al. refer to this as quality control. They ask questions, such as: Is the research question clear and does it match the method used? Does the research design specify the role of the researcher? How will data be collected to maximize the phenomenon of interest and to answer the research question (with specific protocols if multiples data collectors are utilized)? When and how will coding checks, data quality checks, and peer reviews take place? and Do the findings match across sources of data and the developing theory?

External reliability is defined as “whether independent researchers would discover the same phenomena or generate the same constructs in the same or similar setting” (Schensul et al., 1999, p. 275). Miles et al. (2014) consider this issue as one of “relative neutrality and reasonable freedom from unacknowledged researcher biases” (p. 311). They ask questions that relate to the ability to replicate a study. Are the methods and procedures well described, including the sequence for data collection and analysis and the development of conclusions (thus, is there an audit trial)? Are researcher biases addressed? Has the researcher linked conclusions to the data and shown potential hypotheses that were refuted and why? and Are the data available for others to analyze?

GENERALIZABILITY

Generalizability is often referred to as external validity or transferability. Maxwell (1992) defines generalizability as “the extent to which one can extend the account of a particular situation or population to other persons, times, or settings than those directly studied” (p. 293). Generalizability is applicable with the development of theory from qualitative work and the utility of the theory to other persons in similar situations.

Generalizability has typically been a concern of quantitative researchers, whereas many qualitative researchers and texts ignore this issue. There are instances when
generalizability is not a concern for the qualitative researcher. Some qualitative methods are not intended to produce generalizable findings but still generate valuable pieces of work (such as with some phenomenologies, individual case studies, and narrative/storytelling). However, Morse (1999b) points out that for qualitative research to be useable and to advance nursing science, generalizability needs to be addressed. This is particularly true if one of the goals of the qualitative study is to develop theory. Although the means for how qualitatively derived theory is developed is different from quantitatively derived theory, the qualitatively derived theory should still meet strict evaluative criteria (Morse, 1997), and thus, will be applicable beyond the immediate group studied.

ADEQUACY OF THE SAMPLE

Recruitment and sampling can be labor intensive and time consuming, but are critically important to a qualitative project (Thomas, Bloor, & Frankland, 2007). The adequacy of the sample, including both selection of and sample size, will influence the credibility (thus validity, reliability, and generalizability) of qualitative results. There is a general misconception that in qualitative research the sample size is small and homogeneous and many times not an important issue to address when developing a qualitative research proposal. On the other extreme, some researchers overcompensate and plan a study with a huge number of participants to be comparable to a quantitative study. A sample that is too large and diverse may become cumbersome and will not allow for the in-depth analysis that is one of the hallmarks of qualitative research (Sandelowski, 1995). There is no hidden formula that helps the researcher determine sample size, but Morse (1998) has suggested two principles that need to be considered: First, the more units of data generated from each participant the fewer participants needed, and second, the broader the research topic, the larger the sample will need to be in order to reach saturation or redundancy. Here, we consider those and other factors that should be combined to assist the researcher in determining an adequate sample. The factors addressed are the overall purpose of the study, the research question/topic of interest, type of purposive sampling to be employed, the sampling unit (e.g., number of participants versus units of data: number of interviews or incidents), and the research approach (or method) that undergirds the study.

PURPOSE OF THE RESEARCH

The purpose of qualitative research studies range from a study that aims at pure description (such as in a case study, narrative, and focus group) to the development of theory (such as in phenomenology, grounded theory, and ethnography). Thus, sample size would be influenced by the overall purpose of a study. If the purpose of a study is to describe, for example, a person’s life or a specific event/phenomenon, few participants may be needed (for a case study, even a sample of one may be adequate). In this situation, the quality and quantity of information obtained from each individual needs to be significant, so multiple in-depth interviews alone or along with other sources of data, such as journal entries, need to be collected. The purpose of
focus group research also is descriptive in nature; however, the sample size may be larger than with other qualitative methods. This follows Morse’s (1998) principle given earlier, which states that a larger sample is needed when fewer data are obtained from each participant. In focus groups when there are eight to 10 participants, limited or superficial information can be obtained from each participant (Morgan, 1998). Thus, the number of participants needed to adequately address a topic is at least 25 to 30, depending on the topic and homogeneity of the groups. If the purpose of a research project is to develop theory, then the size and diversity of the sample in terms of numbers of participants, number of units of data (interviews/observations), and data sources need to be significant.

RESEARCH QUESTION/TOPIC

The research question or topic of interest will also have a major influence on sample selection and the sample size. One of the tenets of qualitative research is to maximize the phenomenon. The researcher needs to pick the sample that will best answer the research question or examine the topic of interest. The sample is selected based on the data that can be obtained from the participant. Sampling in this manner is biased in that the researcher is looking for participants who are most knowledgeable, thus, have the greatest experience related to the research topic (Morse, 1998). Demographic delimiters, such as gender, ethnicity, socioeconomic factors, and so on, that may be used in quantitative research may hold no analytic value in sample selection strategies for qualitative studies. Keeping this in mind, within an individual study, some of these demographic factors may actually turn out to influence the participants’ experiences. If this does occur, then data for these individual groups need to be collected and, if possible, within the constraints of an individual study, sufficient data need to be collected from each group to reach saturation. A second piece to remember in terms of demographics, particularly age, gender, and ethnicity, is that federal funding guidelines presently mandate the inclusion of adults and children, males and females, and all ethnic groups in research studies. There may be ways to justify the exclusion of certain groups such as children; however, studies may not be funded when gender or ethnic groups are excluded. If the qualitative researcher truly believes that ethnicity is of analytic importance, then the study needs to be designed to address these different groups.

Sample size is also influenced by the research question or topic. A very specific research question would require a smaller sample than a broader topic. For example, investigating the experiences of wives of critically ill patients (thus, only adult patients) would require a smaller sample than a study that explored families’ experiences (which would include both adult and pediatric patients). In the wife study, although the researcher may want to include women with a variety of experiences (e.g., women of different ages whose husbands were critical as a result of different acute, chronic, or traumatic conditions with varying prognoses), the participant group is still fairly discrete. If the study included all family members, the sample would need to be larger to include some of the same variations as in the wife study, but also husbands, children, parents, siblings, and so on. Some of these subgroups may not be dissimilar, so would
not need to be saturated separately; however, until enough data are obtained from these groups, this would not be known.

**TYPE OF SAMPLING**

The type of sampling is another factor to consider when designing a study. Purposive sampling is the mainstay of qualitative research and is defined as the selection of data sources to meet the needs of the study. These data sources may be individuals who have knowledge of or experience with the research topic of interest or other types of data that would enhance the researcher’s ability to comprehensively understand the topic of interest. According to Sandelowski (1995), there are three variations within purposive sampling: demographic, phenomenal, and theoretical. Demographic variation is based on the belief that a demographic factor is analytically important to the study. Participants are selected based on both their experience with the topic of interest and the demographic factor, such as gender. Phenomenal variation incorporates dimensions or characteristics associated with the phenomenon of interest. For example, as described earlier, in the study of wives of critically ill patients, some variations would be the prognosis and cause of the critical illness. Participants would be purposively selected to include these variations. Both demographic and phenomenal sampling decisions are made prior to beginning the study. The third variation is theoretical sampling. Here, sampling decisions are made based on the analytic needs of the study. The researcher may begin by using some demographic or phenomenal variations, but as the study progresses, subsequent sampling is driven by the developing theory or model, therefore, data sources are chosen that have specific characteristics that will expand, validate, or refute the emerging model.

**UNITS OF DATA**

A unit of data is a more useful term to consider when designing a qualitative research study. This gives a clearer idea of the sample than simply estimating in a proposal or documenting in the final report the number of study participants. According to Morse (2015, p. 1170), “all data are not equal.” A unit of data can be an interview, a day of observation, or a single unit of observation (when the incident of interest occurs), or some other data source such as hospital records or newspaper clippings. Many times reviewers are leery when they read that the sample included “only” 10 participants. What is missing in this type of documentation is the fact that each participant was interviewed three times, wrote daily journals entries, or were observed for 8 hours during interactions with multiple family members (so many incidents of interest were observed). When completing a human subject’s protection application, an estimate of the number of participants is needed, but for research funding proposals, the addition of the units of data would strengthen the proposal. So, instead of writing “it is estimated that 15 participants will be needed to reach saturation,” the researcher could say, “it is estimated that 30 interviews will be needed” (so this could mean 15 participants interviewed twice, 10 participants interviewed three times or 10 participants interviewed once, and 10 others interviewed twice). These units of data need to be carefully thought out to ensure that the number of units proposed will be adequate to
reach data saturation, yet not too many that the data are “oversaturated” and extensive analysis becomes difficult.

SAMPLE SIZE BASED ON QUALITATIVE APPROACH

There is no set rule that determines sample size in qualitative research; however, there are some methodological guidelines that may assist the researcher. It must be remembered, that sample size alone is not as important as the number of units of data. So in terms of phenomenology, a small sample of less than 10 participants⁹ is used, however, multiple interactions (i.e., interviews/conversations, writing samples, journals, and other artistic units such as poetry, art, or music) with the participants are needed. A grounded theory approach would require somewhere between 30 and 50 participants and include both interviews and observation (Morse, 1994). An ethnographic approach would also require 30 to 50 participants, along with multiple other data sources (Morse, 1994). The number of participants in focus group research depends on the research question and the diversity of the participants. However, the general rule is that five groups of six to 10 participants are needed (Morgan, 1998). Case study research can be undertaken with one participant, but requires numerous units of data (so multiple interviews or other data sources). Case studies can also be done with an event or other type of incident (this becomes the case). Here, the number of participants and other sources of data will depend on the event chosen and the amount of data needed to fully understand the case (Creswell, 2013). Narrative/storytelling usually uses long and detailed descriptions from one or two individuals. These narratives are a result of multiple in-depth interviews or discussions with the participant(s).

DATA COLLECTION

Throughout the data collection, there are numerous places where threats to validity, reliability, and generalizability can occur. These include the research participant, the researcher as the data collector (specifically during interviewing and observations), and the specific techniques used for data collection.

THE RESEARCH PARTICIPANT

The quality of the collected data depends on the participant. In order for a participant to provide useful data during an interview, the participant must have had the experience that is being investigated, has processed and remembers the experience, and be willing and able to articulate that experience (Nunkoosing, 2005). Not all individuals who agree to participate in a research study will fit those criteria. Some people are unable to relay their experiences. There are many possible reasons for this, including:

- They do not want to tell the story
- They have not processed the experience yet or they are still in the midst of the experience (here observation may be a more appropriate approach)
- They do not remember the event (e.g., as a result of the trauma of the experience or the severity of an illness)
518 VI Measurement Issues

- The experience was too personal
- Fear that relating the experience will get them or someone else in trouble
- The interview is not being conducted in a private place

Other threats to the validity of the collected data include the participants telling the researcher what they think the researcher wants to hear and finding clues while reading the informed consent form (Morse, 2008b) or while listening during an interview and repeating back what they have read or heard (e.g., the researcher uses the word coping and participant picks up on it and uses it). A second threat is acting differently than usual or acting the way they think they should act when being observed. An additional threat to both validity and reliability is the participant providing inconsistent, ambitious, and/or contradictory responses during an interview or observation (Watson, 2006).

There are a number of ways to minimize these threats to validity and reliability. These include: increasing the sample size, performing multiple interviews (this may allow the participant time to process the experience, become comfortable with the interview process and the researcher, and allow time for clarification and/or explanation), completing multiple periods of observation (after a period of time, the influence of being watched will decrease), seeking out negative cases (those individuals who do not support the developing hypotheses or model), and performing member checks.10

THE RESEARCHER AS DATA COLLECTOR: INTERVIEWING

The researcher (or person collecting the data) has a major influence on the data collected and the quality and accuracy of that data (in other words, validity and reliability of the data). As in all research, prior to beginning the study, everyone involved in data collection needs to understand the study and be trained in the techniques of interviewing and/or observation. During the initial set of interviews and periodically throughout the research study, the interviews are critiqued for style (e.g., if the interview is unstructured, does it flow as an unstructured interview should), leading questions, advice giving, interruptions, and researcher biases.

Threats to validity and reliability are inherent in the process of interviewing a research participant, as interviews are not neutral tools for data collection.11 The interview itself is influenced by both the interviewer and interviewee. The interviewer creates the scene and as such is in the position of power, controlling the situation under which an interview takes place, what questions are asked and answers given (Fontana & Frey, 2000; Nunkoosing, 2005). The level of formality between the interviewer and interviewee may also change the flow of the interview and openness of the interviewee’s responses (Hewitt, 2007). During the interview, personal characteristics of the interviewer, such as age, race, gender, and profession, will influence how well the interviewer and interviewee connect, and thus, the data obtained during the interview. How questions are asked and what words12 the interviewer uses, which may be subject to interviewer bias based on personal beliefs and values, also will affect the participants’ responses.
Using a translator during the interview process may be necessary in some cases, however, this poses problems with data validity (Esposito, 2001). Some of the general threats relate to the interpretation of the question by the translator, including: Was the question understood as it was supposed to be? Was it translated as it was meant to be? and Was the question even asked? Other concerns are a result of cultural and language issues (Twinn, 1997). Can the questions or items even be translated or are there words or cultural aspects that do not translate (Lopez, Figueroa, Connor, & Maliski, 2008)? The same concerns apply to the responses given by the participant; that is, are they translatable, and within what context of the conversation. Thus, a literal word-for-word translation may not have true contextual meaning. In order to enhance validity, a competent translator should be trained as an interviewer and be an active part of the research project. He or she needs to understand the purpose of the study, why certain questions are asked, and what probes are needed to adequately address the research topic. If at all possible, the translator should understand not only the language, but also the culture, so that questions can be preevaluated for their relevance to that particular cultural group. In addition, the chosen translator may need to know regional language variations among participants. Interview questions also need to be pretested through this translation process to ensure that valid data can be obtained.

Interviewing is an art and researchers need to critique their own interview style. Within this process threats to validity and reliability can be addressed. Some important things to consider are the amount of talking by the interviewer versus the interviewee. A “good” interview will have enormous blocks of interviewee text with minimal talking by the interviewer. Other things to critique and that are of concern within the interview are:

- Multiple questions being asked at once
- Interrupting the participant before he or she has completed the answer
- Summarizing or verifying what the participant said
- Teaching, preaching, or counseling the participant (you are there as a researcher, not as a nurse, you can refer the participant to the appropriate source for information once the interview is finished)
- Changing directions before the participant has told his or her whole story
- Using labels or technical terms
- Asking value-laden questions or questions based on personal biases
- Asking closed-ended questions (at times, some closed-ended questions are warranted, but for the most part, these should be avoided)

There also are technical problems that need to be preaddressed prior to beginning the process of interviewing. All interviewers need to know how to use the audio-recording equipment (e.g., does the recorder warn you before the battery run outs or the digital recording space is full?), have spare batteries, and an adequate built-in or external microphone, so voice quality will not be compromised if the setting for data collection is less than adequate (always try to do the interview in a quiet setting without distractions). Another problem is loss of data; either the interview did not record or was lost (e.g., erased by accident or not uploaded properly on to the computer.
before being erased). Again, knowing how to properly work the equipment (both the recorder itself and the computer program for uploading and maintaining the digital data) should prevent an interview from not being recorded and uploaded, and keeping a digital copy of the interview on the computer (and one in a backup drive) until a hard copy of the interview is completed will prevent potential loss of data.

**THE RESEARCHER AS DATA COLLECTOR: OBSERVING**

The collection of observational data is more than just a documentation of visual information, but includes observations using all of the researcher’s senses. This technique is used when the information needed to address the research question cannot be obtained by interviews alone and/or when this type of data would augment other types of data (Jorgensen, 1989). Most qualitative observation is done by the researchers (or research assistant) in person, but there are instances when digital video recording may be an option (issues related to video recording are discussed later). One of the fundamental aspects of participant observation (and one that enhances data validity) is that data collection takes place in the natural context in which it occurs and among those individuals who would naturally be part of that context.

There is a range of participation by the researcher who is collecting observational data, from complete observer to a complete participant. Each level has implications in relation to the quality of the data collected. A complete observer (which is similar to video recording) does not interact with the individuals or the setting being observed. This has been described as being “a fly on the wall” (Morse & Field, 1995, p. 109). The disadvantage of being a complete observer is that the researcher’s ability to interview or clarify why a behavior has occurred is not possible. As the researcher becomes more of a participant (either participant-as-observer in which the researcher is part of the setting and acts as a researcher [observer] only part of the time, or observer-as-participant in which the researcher is primarily in the setting to observe, but helps out once in a while [Morse & Field, 1995]), other potential threats to validity and reliability can occur. The more a researcher becomes part of a setting, the less objective (or more bias) the observations become as the researcher may begin to see events through the eyes of the participants. On the other hand, active participation in a setting may facilitate the acquisition of insider knowledge, so it will help the researcher to determine when the event(s) of interest will occur and potentially what and where to observe.

There are numerous threats to both validity and reliability when using participant observation as a technique for data collection. A threat to validity is possible behavioral changes on the part of those being observed when the researcher is present. These changes decrease over time, so increasing the length of time in the setting should enhance the “truth value” of what is being observed. It is also important to be allowed to conduct observations at any time that the researcher deems appropriate so that the individuals being observed do not prepare for the visit (Morse & Field, 1995).

Other threats to validity and reliability include not knowing what is important to observe, thus observing and documenting the wrong information;
missing significant events (because you are observing something else); prematurely determining what is important to observe, so missing the phenomenon of interest; and incomplete or incorrect documentation of the phenomenon in terms of field notes.\textsuperscript{13}

Participant observation rarely consists of solely observational data. Thus, many of the concerns previously raised can be addressed by using multiple techniques for data collection such as interviews and documents. Proper documentation of the phenomenon being observed in a field note (done at regular intervals so information is not forgotten or missed) will also strengthen both the validity and reliability of the collected data. Although the observation itself cannot be critiqued, the field notes from the observatory periods can be evaluated. During this evaluation, questions can be raised as to what has been observed and why, and the comprehensiveness of the field note. This process is extremely important because these field notes are the data that will be used for analysis purposes.

Digital video recording has been suggested as a way to address some of the limitations of participant observation. According to Bottorff (1994), three reasons for using video recording are: “when behaviors of interest are of very short duration, the distinctive character of events change moment by moment, or more detailed and precise descriptions of behaviors and/or processes than possible with ordinary participant observation are required” (p. 244). The two advantages of video recording are density and permanence (Bottorff, 1994). Density refers to the amount of information that can be simultaneously recorded. Density is significantly higher with video recording versus participant observation; however, as Bottorff (1994) points out, there are also limitations to this technique. The camera is still only picking up certain events, depending on where it is aimed (so it may pick up what’s going on in the room, but miss closer facial expressions), microphones may pick up extraneous noises and not the vocalizations of importance, and, as with participant observation, behaviors may change because of the camera. Permanence refers to the ability to have the event recorded allowing the researcher to review the event as often as needed and the analysis can be demonstrated to others. Having a video recording of an event definitely increases data validity and reliability; however, in some situations, the inability to have data beyond the recording (similar to complete observer) is present. So, as different hypotheses are developed, unless data collection is ongoing, testing will not be possible. Thus, other forms of data collection, such as is available in a simulation lab where other data can be captured simultaneously, would be a way to address such issues.

**TECHNIQUES FOR DATA COLLECTION**

One of the most common mistakes made by researchers is not using the appropriate techniques for data collection to adequately investigate the research topic of interest. Three areas to consider are sources of data, group versus individual interviews, and in-person versus telephone versus online interviews. In order to have valid data that truly reflects reality, multiple sources of data are usually required. Simply interviewing participants may only give a partial picture, because they may not be able to relay the whole story (there are many things that are done unconsciously, so cannot be easily verbalized). Thus, in many instances, participant observation and other sources
of data, such as medical records, are also needed. There are, however, other times when the researcher is interested solely in past experiences and participant observation is not needed and would provide useless data. Internet sources of data may provide a larger, more diverse sample, yet may also be unfairly biased toward those individuals who are literate and have access to a computer (Hamilton & Bowers, 2006; Hunt & McHale, 2007). Using Internet chat rooms may provide participants the ability to freely speak and have new thoughts and ideas brought out through a group discussion (similar to participation in a focus group). However, once again, the sample cannot easily be theoretically derived because the participants are self-selected.

Another common error is the use of group interviews or focus groups when in-depth individual interviews are needed to sufficiently explore the topic. Although the group interaction within a group interview or focus group may stimulate discussion, there is always the possibly of group think and the inability for each participant to tell his or her whole story either because of time constraints or fear of presenting an opposing view or revealing a personal truth in this type of interview. The use of in-person versus telephone versus online interviews is another aspect to consider. In-person interviews may provide a connection between the researcher and participant that will allow for more open discussion and also permit the researcher to observe the participant's responses during the interaction. Phone interviews are many times used for the initial interview when the participant is not easily accessible or more commonly used for follow-up interviews (Sturges & Hanrahan, 2004). Telephone interviews as the sole method of data collection will hide many of the personal characteristics of the interviewer that may subtly influence the interview, but the lack of the personal connection between the interviewer and interviewee may impede the openness of the participant. However, interviewees also may be more willing to say things that they might be more hesitant to say during a face-to-face interview (Novick, 2008). Internet-based interviewing provides an additional means to collect data. Chat room discussions are one means to collect data focused on a particular topic that is posed by the researcher or as part of a discussion group (e.g., related to a certain health-related issue). E-mail interviews are another technique for interviewing. This format allows the interviewee time to thoughtfully answer questions, and the freedom to say things that he or she may not be willing to say in person (Hunt & McHale, 2007; Ratislavová & Ratislav, 2014). However, the response may be restrained by a participant's inability or unwillingness to elaborate or write long, detailed responses (Hamilton & Bowers, 2006). There is also an issue with trustworthiness of the data; because the sample may be self-elected, the researcher may not know for sure with whom he or she is communicating (Hunt & McHale). Finally, because the process takes time, interviewees may lose focus leading to incomplete data sets and a high rate of attrition.

Finally, the use of wrong interview structure can pose a serious threat to validity and reliability. Using semistructured or structured interviews too early in a research study when little information is known about the phenomenon of interest can result in the true phenomenon not emerging. Many times the researcher develops a list of structured interview questions that presuppose which concepts are of importance to the participant, thus, not allowing these to emerge from the data. On the other hand,
if the research is at the point of hypothesis testing and beginning theory development, unstructured interviews will not provide the data needed for this to occur.

**DATA ANALYSIS**

During the process of data analysis there also are numerous places where threats to validity and reliability can occur. But before data analysis is discussed, data quality needs to be addressed. Of number one importance to data quality is the accuracy of the transcribed interviews and field notes. Accuracy of the data set is the responsibility of the researcher. Each transcribed piece of data needs to be checked for accuracy because simple things like inaccurate punctuation or a mistyped word can change the entire meaning of a sentence (Gray, 2014). Transcriptionists need to be spot checked to ensure their accuracy in transcribing: Do they understand the terminology, are they just transcribing *words* versus words within context, and is there language or cultural-specific content that is being misunderstood (MacLean, Meyer, & Estable, 2004)? In addition, the researcher needs to fill in gaps in the transcription (spots where the typist could not understand or hear the voice). The researcher who actually performed the interview is the best person to do this because he or she may be able to reconstruct the interview.

There are areas that researchers should consider when having interviews transcribed. These include which features of the interview are important and need to be part of the transcript versus what can be ignored (Sandelowski, 1994). For example, how will emotional expression be described? Are pauses important and how will you and the typist determine how to document them? and Does every “uh huh” need to be included? Each of these decisions should be made based on the overall purpose of the research and the method chosen for the study.

The validity and reliability of a study is integrally tied to the process of data analysis. The findings and conclusions of a study need to be cohesively linked to the data. There are key areas that should be addressed during the planning stage of a research project, so that validity and reliability will not be compromised once analysis begins. First, the choice of type of analysis needs to match the overall purpose of the study, the methods chosen, and the type of data collected (for further information on analysis, see Guest, MacQueen, & Namey, 2012). Thus, a constant comparative analysis will not work for a focus group study, but is the technique of choice for a grounded theory. Content analysis is facilitated when every participant is asked and responds to the same question. However, if unstructured open-ended interviews are done, a different form of content analysis may be applied (see Hsieh & Shannon, 2005). Second, preplanning how model verification, coding checks, and audit trails will occur will make the process proceed smoothly. Finally, assembling a team to assist with the analysis will help address researcher bias (especially if the team is not all nurses or at least not all nurses from the same specialty area).

Computer-assisted qualitative data analysis software (CAQDAS) is a wonderful asset for electronically storing large data sets, searching and retrieving qualitative data, and as an aid for the researcher with analysis (Wickham & Woods, 2005). The strength of the analysis is dependent, not on the program chosen, but on the
quality of the data collected and analytic decisions made by the researcher along with the use of well-established qualitative analysis strategies (Thompson, 2002). There are a number of programs available, including but not limited to: Ethnograph (www.qualisresearch.com), ATLAS.ti (www.atlasti.com), MAXQDA (www.maxqda.com), HyperRESEARCH (www.researchware.com), and QSRNVivo (www.qsrinternational.com). Each program has both positive and negative aspects, thus choice of a particular program is usually determined by access and support within an institution or research team. All of these programs have a huge learning curve, but once the researcher and/or research team masters the program, the program allows for group analysis sessions in which the analysis decisions can be shared with other members of the team. In addition, supporting data and negative cases can be easily found to validate the assumptions that are being posed by the researcher. One main concern with the use of these programs and a major threat to validity is the belief by novice researchers that the programs actually analyze the data (MacMillan & Koenig, 2004). At times, the researcher begins to work deductively, having the program search for specific words or phrases throughout the data set. There are points during analysis at which this tool is helpful, however, the researcher needs to be careful that the phrase, first, is important, and, second, that it is consistently used throughout the data set. The major threat is that a few participants will use a word or phrase numerous times, so the number of uses seems relevant, but others may never use the term.

Another problem that frequently arises during analysis includes premature closure of analysis and development of categories/models. Here, either researcher bias comes into play or the researcher observes a pattern of some type early on in the analysis process and decides that “this is it.” When this occurs, data are inaccurately placed into categories (here, the researcher begins to work deductively instead of inductively), thus, when the real pattern or categories appear, they are missed. For a beginning researcher, working with an experienced qualitative researcher and/or a research team should prevent this from occurring. The research team needs to independently code the data and then compare codes and developing categories. Areas of incongruence need to be discussed using specific examples from the data set. If incongruencies continue, further data collection may be needed to examine each area. The experienced researcher should be able to question the research team to ensure that the categories or conclusions are grounded in the data and ask the “what about” questions. The team needs to have rich description from the data to support developing categories to answer questions, such as “you say that this works for situation x, what about in situation y?“ (e.g., once trust is established, wives of critically ill patients are able to relinquish care to the nursing staff; does the same hold true for parents?). This back and forth and potential verification with the original participants or secondary informants will help to strengthen the analysis.

**SUMMARY**

Qualitative nurse researchers have made significant and sustained contributions to nursing science and, in the process, several nurses have contributed to the methodological literature as well. Given the complexity of issues of interest to nursing, qualitative methods provide valuable tools for the nurse researcher. Yet, these approaches
demand careful attention to the paradigmatic perspectives undergirding a particular method or study. Potential threats to the integrity of qualitative research projects emerge in all phases of the research, from planning the study through data collection, analysis, and reporting the findings.

Qualitative researchers cannot simply avoid issues surrounding the notions of validity, reliability, or generalizability. It is imperative that threats are carefully analyzed and strategies to reduce the threats are integrated into the design and clearly articulated for grant reviewers. As the study evolves, the researcher must be vigilant for new threats or shifts from the original strategies. Finally, in dissemination, limitations should be carefully addressed to guide subsequent research.

NOTES

1. For a deeper discussion of techniques for evaluating the state of the science and designing subsequent projects to advance the science, see Hupcey and Penrod (2005) and Penrod and Hupcey (2005a, 2005b).
2. If an inappropriate method is chosen, then the research question cannot be answered, for example, trying to investigate an extremely personal issue in a focus group may result in little or false information being given by the participant.
3. For a comprehensive discussion of levels of qualitatively derived theory, see Morse (1997).
4. See Morgan (1998) for greater detail on numbers of groups and sample size.
5. Notice here we say “project” and not “study,” because although an individual study may begin the process of theory development, numerous studies are needed, thus, the word project, to build solid high-level theory (Hupcey & Penrod, 2003).
6. Many small studies do not have the time or resources to collect enough data on each group when this occurs; in these situations, the researcher must acknowledge this as a limitation of the study and present this as a future avenue for research.
7. Sample size in qualitative research is always estimated, for the researcher does not know a priori the number of sampling units or participants that will be needed to reach analytic saturation. However, the researcher must have a theoretically based justification for this estimate.
8. When participants are interviewed a varying number of times, the reason why this is anticipated (or occurred) needs to be explained and, at times, justified.
9. According to Morse (1994), approximately six participants are needed to understand the essence of an experience.
10. Member checks are purported to be a way to help enhance the credibility of the analysis, by adding evidence that the interpretations of the researchers represent reality. “Participants can act as judges, evaluating the major finding of a study” (Miles et al., 2014, p. 309).
11. For a complete discussion of the types of interviews, see Waltz, Strickland, and Lenz (2010, pp. 287–300).
12. Spoken words are always ambiguous and may be interpreted differently by the interviewer and participant (Fontana & Frey, 2000).
13. For a complete description of field notes, see Morse and Field (1995, pp. 111–115).
VI ■ Measurement Issues

REFERENCES


Morse, J. M. (2008a). Expanding the application of qualitative findings. *Qualitative Health Research, 18*, 1459–1460.


© Springer Publishing Company