

Evaluation and Testing in

NURSING EDUCATION

Marilyn H. Oermann
Kathleen B. Gaberson



EVALUATION AND TESTING IN NURSING EDUCATION

Marilyn H. Oermann, PhD, RN, ANEF, FAAN, is the Thelma M. Ingles Professor of Nursing at Duke University School of Nursing, Durham, North Carolina. She is an author or coauthor of 21 books and many articles on evaluation, teaching in nursing, and writing for publication as a nurse educator. She is the Editor-in-Chief of *Nurse Educator* and the *Journal of Nursing Care Quality* and past editor of the *Annual Review of Nursing Education*. Dr. Oermann lectures widely on teaching and evaluation in nursing.

Kathleen B. Gaberson, PhD, RN, CNOR, CNE, ANEF, is the owner of and principal nursing education consultant for OWK Consulting, Pittsburgh, Pennsylvania. She has over 35 years of teaching and administrative experience in graduate and undergraduate nursing programs. She is a coauthor of 10 nursing education books and an author or coauthor of numerous articles on nursing education and perioperative nursing topics. Dr. Gaberson presents and consults extensively on nursing curriculum revision, assessment and evaluation, and teaching methods. Dr. Gaberson is the Associate Editor for Research of the *AORN Journal* and also serves on the *Journal* Editorial Board.



EVALUATION AND TESTING IN NURSING EDUCATION

Sixth Edition

Marilyn H. Oermann, PhD, RN, ANEF, FAAN
Kathleen B. Gaberson, PhD, RN, CNOR, CNE, ANEF

Copyright © 2021 Springer Publishing Company, LLC
All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior permission of Springer Publishing Company, LLC, or authorization through payment of the appropriate fees to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400, fax 978-646-8600, info@copyright.com or on the Web at www.copyright.com.

Springer Publishing Company, LLC
11 West 42nd Street
New York, NY 10036
www.springerpub.com
<http://connect.springerpub.com/home>

Acquisitions Editor: Adrienne Brigido
Compositor: diacriTech

ISBN: 978-0-8261-3574-2
ebook ISBN: 978-0-8261-3575-9
Instructor's Manual ISBN: 978-0-8261-3576-6
Instructor's PowerPoints ISBN: 978-0-8261-3577-3
DOI: 10.1891/9780826135759

Qualified instructors may request supplements by emailing textbook@springerpub.com

19 20 21 22 23 / 5 4 3 2 1

The author and the publisher of this Work have made every effort to use sources believed to be reliable to provide information that is accurate and compatible with the standards generally accepted at the time of publication. The author and publisher shall not be liable for any special, consequential, or exemplary damages resulting, in whole or in part, from the readers' use of, or reliance on, the information contained in this book. The publisher has no responsibility for the persistence or accuracy of URLs for external or third-party Internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

Library of Congress Cataloging-in-Publication Data

Library of Congress Control Number: 2019917090

Contact us to receive discount rates on bulk purchases. We can also customize our books to meet your needs.
For more information please contact: sales@springerpub.com

***Publisher's Note:* New and used products purchased from third-party sellers are not guaranteed for quality, authenticity, or access to any included digital components.**

Printed in the United States of America.

CONTENTS

Preface *vii*

Instructor's Resources *xiii*

■ **PART I. Concepts of Assessment**

1. Assessment and the Educational Process **3**
2. Qualities of Effective Assessment Procedures: Validity, Reliability, and Usability **23**

■ **PART II. Testing and Other Assessment Methods**

3. Planning for Testing **47**
4. True–False and Matching **69**
5. Multiple-Choice and Multiple-Response **79**
6. Short-Answer (Fill-in-the-Blank) and Essay **101**
7. Assessment of Higher Level Learning **119**
8. Test Construction and Preparation of Students for NCLEX® and Certification Examinations **141**
9. Assessment of Written Assignments **159**

■ **PART III. Test Construction and Analysis**

10. Assembling, Administering, and Scoring Tests **177**
11. Testing and Evaluation in Online Courses and Programs **199**
12. Test and Item Analysis: Interpreting Test Results **221**

■ PART IV. Clinical Evaluation

- 13. Clinical Evaluation Process 253
- 14. Clinical Evaluation Methods 267
- 15. Simulation and Objective Structured Clinical Examination for Assessment 297

■ PART V. Issues Related to Testing and Evaluation in Nursing Education

- 16. Social, Ethical, and Legal Issues 311
- 17. Grading 329
- 18. Program Evaluation and Accreditation 351

■ Appendices

- Appendix A** Quick Reference Guide for Writing Different Types of Test Items With Examples 377
- Appendix B** Testing Resources for Nurse Educators 389
- Appendix C** Code of Fair Testing Practices in Education 391
- Appendix D** National League for Nursing Fair Testing Guidelines for Nursing Education 399
- Appendix E** Standards for Teacher Competence in Educational Assessment of Students 403

Index 405



PREFACE

All teachers at some time or another need to assess learning. The teacher may write test items; prepare tests and analyze their results; develop rating scales and clinical evaluation methods; and plan other strategies for assessing learning in the classroom, clinical practice, online courses, simulation, and other settings. Often teachers are not prepared to carry out these tasks as part of their instructional role. This sixth edition of *Evaluation and Testing in Nursing Education* is a resource for teachers in nursing education programs and healthcare agencies; a textbook for graduate students preparing for their role as a nurse educator; a guide for nurses in clinical practice who teach others and are responsible for evaluating their learning and performance; and a resource for other healthcare professionals involved in assessment, measurement, testing, and evaluation. Although the examples of test items and other types of assessment methods provided in this book are nursing oriented, they are easily adapted to assessment in other health fields.

The purposes of this book are to describe concepts of assessment, testing, and evaluation in nursing education and prepare teachers for carrying these out as part of their roles. The book presents qualities of effective assessment procedures (reliability, validity, and usability); how to plan for testing, assemble and administer tests, and score tests; how to write all types of test items and develop assessment methods; strategies for assessing higher level learning; and testing and evaluation in online courses and programs. The book describes the evaluation of written assignments in nursing, the development of rubrics, clinical evaluation, methods for evaluating clinical performance, and using simulation and objective structured clinical examinations (OSCEs) for evaluation. A new chapter prepares educators to analyze the performance of the test as a whole and of individual test items and to interpret test scores; this chapter includes many examples and exhibits to help readers understand these analyses and make informed decisions about their tests. This edition also examines the social, ethical, and legal issues associated with testing and evaluation in nursing; grading; and program evaluation (with a new section on accreditation of nursing education programs). The Appendices provide a quick reference guide for writing different types of test items (with examples) and

other testing resources. The content is useful for teachers in any setting who are involved in evaluating others, whether they are students, nurses, or other types of healthcare personnel.

Chapter 1 addresses the purposes of assessment, testing, measurement, and evaluation in nursing education. Differences between formative and summative evaluation and between norm-referenced and criterion-referenced measurements are explored. Because effective assessment requires a clear description of *what* and *how* to assess, the chapter describes the use of outcomes for developing test items, provides examples of outcomes at different taxonomic levels, and describes how test items would be developed at each of these levels.

In Chapter 2, qualities of effective assessment procedures are discussed. The concept of assessment validity, the role of reliability, and their effects on the interpretive quality of assessment results are described. Tests and other assessment instruments yield scores that teachers use to make inferences about how much learners know or what they can do. *Validity* is the adequacy and appropriateness of those interpretations about learners' knowledge or ability based on those scores. Current ways of thinking about reliability and its relationship to validity are explained. Also discussed in Chapter 2 are important practical considerations that might affect the choice or development of tests and other instruments.

Chapter 3 describes the steps involved in planning for test construction, enabling the teacher to make good decisions about what and when to test, test length, difficulty of test items, item formats, and scoring procedures. An important focus of the chapter is how to develop a test blueprint and then use it for writing test items; examples are provided to clarify this process for the reader. Broad principles important in developing test items, regardless of the specific type, are described in the chapter.

There are different ways of classifying test items: One way is to group them according to how they are scored—objectively or subjectively. Another way is to group them by the type of response required of the test-taker—selected or constructed response—which is how we organized the chapters. Selected-response items require the test-taker to select the correct or best answer from options provided by the teacher. These items include true–false, matching, multiple choice, and multiple response. Constructed-response items (fill-in-the-blank and essay) require the test-taker to supply an answer rather than choose from options already provided. Chapters 4 to 6 discuss these test items.

A true–false item consists of a statement that the student judges as true or false. In some forms, students also correct the response or supply a rationale as to why the statement is true or false. True–false items are most effective for recall of facts and specific information but may also be used to test the student's comprehension of the content. Chapter 4 describes how to construct true–false items and different variations, for example, correcting false statements or providing a rationale for the response, which allows the teacher to assess if the learner understands the content. Chapter 4 also explains how to develop matching exercises. These consist of two

parallel columns in which students match terms, phrases, sentences, or numbers from one column to the other. Principles for writing each type of item are presented, accompanied by sample items.

In Chapter 5, the focus is on writing multiple-choice and multiple-response items. Multiple-choice items, with one correct answer, are used widely in nursing and other fields. This format of test item includes an incomplete statement or question, followed by a list of options that complete the statement or answer the question. Multiple-response items are designed similarly, although more than one answer may be correct. There are three parts in a multiple-choice item, each with its own set of principles for development: (a) stem, (b) answer, and (c) distractors. In Chapter 5, we discuss how to write each of these parts and provide many examples. We also describe principles for writing multiple-response items, including the format used on the NCLEX® (National Council Licensure Examination).

Short-answer (fill-in-the-blank) items can be answered by a word, phrase, or number. One format presents a question that students answer in a few words or phrases. With the other format, completion or fill-in-the-blank, students are given an incomplete sentence that they complete by inserting a word or words in the blank space. On the NCLEX, candidates may be asked to perform a calculation and type in the number or to put a list of responses in proper order. In Chapter 6, we describe how to write different formats of short-answer items. We also explain how to develop and score essay items. Essay items provide an opportunity for students to select content to discuss, present ideas in their own words, and develop an original and creative response to a question. We provide an extensive discussion on scoring essay responses and on developing rubrics.

With higher level thinking, students apply concepts and other forms of knowledge to new situations, use that knowledge to solve patient and other types of problems, and arrive at rational and well thought-out decisions about actions to take. The main principle in assessing higher level learning is to develop test items and other assessment methods that require students to apply knowledge and skills in a *new* situation; the teacher can then assess whether the students are able to use what they have learned in a different context. Chapter 7 presents strategies for assessing higher levels of learning in nursing. Context-dependent item sets or interpretive exercises are discussed as one format of testing appropriate for assessing higher level cognitive skills. Suggestions for developing these are presented in the chapter, including examples of different items. Other methods for assessing cognitive skills in nursing also are presented in this chapter: cases, case studies, unfolding cases, discussions using higher level questioning, debates, video clips, and short written assignments.

Chapter 8 focuses on developing test items that prepare students for licensure and certification examinations. The chapter begins with an explanation of the NCLEX test plans and their implications for nurse educators. Examples are provided of items written at different cognitive levels, thereby avoiding tests that focus only on recall

and memorization of facts. The chapter also describes how to write questions about clinical practice or the nursing process and provides sample stems for use with those items. The types of items presented in the chapter are similar to those found on the NCLEX and many certification tests. When teachers incorporate these items on tests in nursing courses, students acquire experience with this type of testing as they progress through the program, preparing them for taking licensure and certification examinations as graduates.

Through papers and other written assignments, students develop an understanding of the content they are writing about. Written assignments with feedback from the teacher also help students improve their writing ability, an important outcome in any nursing program from the beginning level through graduate study. Chapter 9 provides guidelines for assessing formal papers and other written assignments in nursing courses. The chapter includes criteria for assessing the quality of papers, an example of a scoring rubric, and suggestions for assessing and grading written assignments.

Chapter 10 explains how to assemble, administer, and score a test. In addition to preparing a test blueprint and skillful construction of test items, the final appearance of the test and the way in which it is administered can affect the validity of its results. In Chapter 10, test design rules are described; suggestions for reproducing the test, maintaining test security, administering it, and preventing cheating are presented in this chapter as well.

Online education in nursing continues to expand at a rapid pace. Chapter 11 discusses assessment of learning in online courses, including testing and evaluating course assignments. The chapter begins with a discussion of online testing. To deter cheating and promote academic integrity, faculty members can use a variety of both low- and high-technology solutions. Providing timely and substantive feedback to students is critical in online courses, and we have included a sample rubric for an online discussion forum assignment and evaluation. Clinical evaluation of students in online courses and programs presents challenges to faculty members and program administrators. The chapter includes discussion of methods for evaluating students' clinical performance in an online course. Other sections of this chapter examine assessment of online courses, student evaluation of teaching, and evaluating the quality of online nursing programs.

After administering the test, the teacher needs to score it, interpret the results, and then use the results to make informed decisions. Chapter 12 discusses the processes of obtaining scores, performing test and item analysis, and interpreting the results (for both teacher-made and standardized tests). The chapter examines test score distributions, describes measures of central tendency and variability, and explains their use in interpreting test scores. How to interpret the difficulty index and discrimination index and analyze each distractor are described. Examples of item analyses are provided for true-false, matching, multiple-choice, and multiple-response

items. Exhibits in the chapter illustrate these analyses so readers understand how to analyze and interpret the performance of individual test items. It also suggests ways in which teachers can use posttest discussions to contribute to student learning and seek student feedback that can lead to test-item improvement. Teachers often debate the merits of adjusting test scores by eliminating items or adding points to compensate for real or perceived deficiencies in test construction or performance. We discuss this in the chapter and provide guidelines for faculty in making these decisions. A section of the chapter also presents suggestions and examples of developing a test-item bank. Many publishers also offer test-item banks that relate to the content contained in their textbooks; we discuss why faculty members need to be cautious about using these items for their own examinations.

Chapter 13 describes the process of clinical evaluation in nursing. It begins with a discussion of the outcomes of clinical practice in nursing programs and then presents essential concepts underlying clinical evaluation. In this chapter, we discuss fairness in evaluation, how to build feedback into the evaluation process, and how to determine *what* to evaluate in clinical courses.

Chapter 14 builds on concepts of clinical evaluation examined in the preceding chapter. Many evaluation methods are available for assessing competencies in clinical practice. We discuss observation and recording observations in notes about performance, checklists, and rating scales; written assignments useful for clinical evaluation such as journals, concept maps, case analyses, and short papers; electronic portfolio assessment and how to set up a portfolio system for clinical evaluation; conferences; and other methods such as group projects and self-evaluation. The chapter includes a sample form for evaluating student participation in clinical conferences and a rubric to use for peer evaluation of participation in group projects.

Simulation is used widely for instruction in nursing, and it also can be used for assessment. A simulation can be developed for students to demonstrate procedures and technologies, analyze data, and make decisions. Students can care for the patient individually or as a team. Student performance in these simulations can be assessed to provide feedback or to verify competencies. Some simulations incorporate standardized patients, actors who portray the role of a patient with a specific diagnosis or condition. Another method for evaluating skills and clinical competencies of nursing students is OSCE. In an OSCE, students rotate through stations where they complete an activity or perform a skill, which then can be evaluated. Chapter 15 describes these methods for assessing clinical competencies of students.

Chapter 16 explores social, ethical, and legal issues associated with testing and evaluation. Social issues such as test bias, grade inflation, effects of testing on self-esteem, and test anxiety are discussed. Ethical issues include privacy and access to test results. By understanding and applying codes for the responsible and ethical use of tests, teachers can ensure the proper use of assessment procedures and the valid interpretation of test results. We also discuss selected legal issues associated with testing.

Grading is the use of symbols, such as the letters A through F or pass–fail, to report student achievement. Grading is used for summative purposes, indicating how well the student met the outcomes of the course and clinical practicum. To represent valid judgments about student achievement, grades should be based on sound evaluation practices, reliable test results, and multiple assessment methods. Chapter 17 examines the uses of grades in nursing programs, types of grading systems, how to select a grading framework, and how to calculate grades with each of these frameworks. We also discuss grading clinical practice, using pass–fail and other systems for grading, and provide guidelines for the teacher to follow when students are on the verge of failing a clinical practicum. We also discuss learning contracts and provide an example of one.

Program evaluation is the process of judging the worth or value of an educational program. With the demand for high-quality programs, there has been a greater emphasis on systematic and ongoing program evaluation. Thus, Chapter 18 presents an overview of program evaluation models and discusses evaluation of selected program components, including curriculum, courses, and teaching–learning activities. The chapter includes content on accreditation, accrediting agencies in nursing, types of accreditation, and evaluation of distance education programs. We discuss the development of a systematic plan for evaluation and include a sample format. Student ratings of courses and teachers carry much weight in many schools of nursing; we discuss these ratings and related issues and examine other sources of information about teaching effectiveness.

In addition to this book, we have provided an Instructor's Manual that includes a sample course syllabus, chapter-based PowerPoint presentations, and ready-to-use modules for an online course (with chapter summaries, student learning activities, discussion forum questions, and assessment strategies).

We wish to acknowledge Adrienne Brigido, Director of Nurse Education for Springer Publishing Company, for her enthusiasm and continued support. We also thank Springer Publishing Company for its support of nursing education and for publishing our books for many years.

*Marilyn H. Oermann
Kathleen B. Gaberson*



INSTRUCTOR'S RESOURCES

Evaluation and Testing in Nursing Education, Sixth Edition, includes a robust ancillary package. Qualified instructors may obtain access to ancillary materials by emailing textbook@springerpub.com. Available resources:

- Instructor's Manual:
 - Sample Course Syllabus
 - Chapter Summaries
 - Student Learning Activities
 - Discussion Questions
 - Assessment Strategies
- Chapter-Based PowerPoint Presentations



CONCEPTS OF ASSESSMENT

ASSESSMENT AND THE EDUCATIONAL PROCESS

In all areas of nursing education and practice, assessment is important to obtain information about student learning, evaluate competencies and clinical performance, and arrive at other decisions about students and nurses. Assessment is integral to monitoring the quality of educational and healthcare programs. By evaluating outcomes achieved by students, graduates, and patients, the effectiveness of programs can be measured and decisions can be made about needed improvements.

Assessment provides a means of ensuring accountability for the quality of education and services provided. Nurses, like other healthcare professionals, are accountable to their patients and society in general for meeting patients' health needs. Along the same lines, nurse educators are accountable for the quality of teaching provided to learners, outcomes achieved, and overall effectiveness of educational programs. Educational institutions also are accountable to their governing bodies and society in terms of educating graduates for present and future roles. Through assessment, nurse educators and other healthcare professionals collect information for evaluating the quality of their teaching and programs as well as documenting outcomes for others to review. All educators, regardless of the setting, need to be knowledgeable about assessment, testing, measurement, and evaluation.

■ Assessment

Educational assessment involves collecting information to make decisions about learners, programs, and educational policies. Mislevy (2017) defined *assessment* as gathering information about what students know and can do. Are students learning the important concepts in the course and developing the clinical competencies? With information collected through assessment, the teacher can determine relevant learning activities to meet students' learning needs and help them improve performance. Assessment that provides information about learning needs is diagnostic; teachers use that information to decide on the appropriate content, learning activities, and practice opportunities for students to meet the desired learning outcomes.

Assessment also generates feedback for students, which is particularly important in clinical practice as students develop their competencies and learn to think through complex clinical situations. Feedback from assessment similarly informs the teacher and provides data for deciding how best to teach certain content and skills; in this way, assessment enables teachers to improve their educational practices and how they teach students.

Another important purpose of assessment is to provide valid and reliable data for determining students' grades. Although nurse educators continually assess students' progress in meeting the outcomes of learning and developing the clinical competencies, they also need to measure students' achievement in the course. Grades serve that purpose. Assessment strategies provide the data for faculty to determine whether students achieved the outcomes and developed the essential clinical competencies. Grades are symbols—for instance, the letters A through F—for reporting student achievement.

Assessment generates information for decisions about courses, the curriculum, and the nursing program. In this context, assessment is the process of collecting information for program evaluation and accreditation. Other uses of assessment information are to select students for admission to an educational institution and a nursing program and place students in appropriate courses. A broad view of assessment is that it encompasses the entire process of evaluating learners and institutional effectiveness (Banta & Palomba, 2014).

There are many assessment strategies that teachers can use to obtain information about students' learning and performance. These methods include tests that can be developed with different types of items, papers, other written assignments, projects, small-group activities, oral presentations, e-portfolios, observations of performance, simulation-based assessments, objective structured clinical examinations (OCSEs), and conferences, among others. Each of those assessment strategies as well as others is presented in this book.

Brookhart and Nitko (2019) identified five guidelines for effective assessment. These guidelines should be considered when deciding on the assessment strategy and its implementation in the classroom, online course, skills or simulation laboratory, or clinical setting.

1. *Identify the learning objectives (outcomes or competencies) to be assessed.* These provide the basis for the assessment: The teacher determines whether students are meeting or have met the outcomes and competencies. The clearer the teacher is about *what* to assess, the more effective will be the assessment.
2. *Match the assessment strategy to the learning goal.* The assessment strategy needs to provide information about the particular outcome or competency being assessed. If the outcome relates to analyzing issues in the care of patients with chronic pain, a true–false item about a pain medication would not be appropriate. An essay item, however, in which students analyze a

scenario about an adult with chronic pain and propose multiple approaches for pain management would provide relevant information for deciding whether students achieved that outcome.

3. *Meet the students' needs.* Students should be clear about what is expected of them. The assessment strategies, in turn, should provide feedback to students about their progress and achievement in demonstrating those expectations, and should guide the teacher in determining the instruction needed to improve performance.
4. *Use multiple assessment strategies and indicators of performance for each outcome.* It is unlikely that one assessment strategy will provide sufficient information about achievement of the outcomes. A test that contains mainly recall items will not provide information on students' ability to apply concepts to practice or analyze clinical situations. The extent and depth of student learning is often difficult to measure on a test. In most courses, multiple assessment strategies are needed to determine whether the outcomes were met.
5. *Keep in mind the limitations of assessment when interpreting the results.* One test, one paper, one observation in clinical practice, or one simulation activity may not be a true measure of the student's learning and performance. Many factors can influence the assessment, particularly in the clinical setting, and the information collected in the assessment is only a sample of the student's overall achievement and performance.

■ Tests

A test is a set of items to which students respond in written or oral form, typically during a fixed period of time. Brookhart and Nitko (2019) defined a *test* as an instrument or a systematic procedure for describing one or more characteristics of a student. Tests are typically scored based on the number or percentage of answers that are correct and are administered similarly to all students. Although students often dread tests, information from tests enables faculty to make important decisions about students.

Tests are used frequently as an assessment strategy. They can be used at the beginning of a course or instructional unit to determine whether students have the prerequisite knowledge for achieving the outcomes or whether they have already met them. With courses that are competency based, students can then progress to the next area of instruction. Test results also indicate gaps in learning and performance that should be addressed first. With that information, teachers can better plan their instruction. Tests can be used during the instruction to provide the basis for formative assessment (Miller, Linn, & Gronlund, 2013). This form of assessment is to monitor learning

progress, provide feedback to students, and suggest additional learning activities as needed. When teachers are working with large groups of students, it is difficult to gear the instruction to meet each student's needs. However, diagnostic quizzes and tests reveal content areas in which individual learners may lack knowledge. Not only do the test results guide the teacher in suggesting remedial learning activities, but they also serve as feedback to students about their learning needs. In some nursing programs, students take commercially available tests as they progress through the curriculum to identify gaps in their learning and prepare them for taking the National Council Licensure Examinations, the NCLEX-RN® or NCLEX-PN®.

Tests are used for selecting students for admission to higher education settings and to nursing programs. Admission tests provide norms that allow comparison of the applicant's performance with that of other applicants. Tests also may be used to place students into appropriate courses. Placement tests, taken after students have been admitted, provide data for determining which courses they should complete in their programs of study. For example, a diagnostic test of statistics may determine whether a nursing student is required to take a statistics course prior to beginning graduate study.

By reviewing test results, teachers can identify content areas that students learned and did not learn in a course. With this information, faculty can modify the instruction to better meet student learning needs in future courses. Last, testing may be an integral part of the curriculum and program evaluation in a nursing education program. Students may complete tests to measure program outcomes rather than to document what was learned in a course. Test results for this purpose often suggest areas of the curriculum for revision and may be used for accreditation reports.

■ Measurement

Measurement is the process of assigning numbers to represent student achievement or performance, for instance, answering 85 out of 100 items correctly on a test. The numbers or scores indicate the degree to which a learner possesses a certain characteristic. Measurement is important for reporting the achievement of learners on nursing and other tests, but not all outcomes important in nursing practice can be measured by testing. Many outcomes are evaluated qualitatively through other means, such as observations of performance in clinical practice or simulation.

Although measurement involves assigning numbers to reflect learning, these numbers in and of themselves have no meaning. Scoring 15 on a test means nothing unless it is referenced or compared with other students' scores or to a predetermined standard. Perhaps 15 was the highest or lowest score on the test, compared with other students. Or the student might have set a personal goal of achieving 15 on the test; thus, meeting this goal is more important than how others scored on the test.

Another interpretation is that a score of 15 might be the standard expected of this particular group of learners. To interpret the score and give it meaning, having a reference point with which to compare a particular test score is essential.

In clinical practice, how does a learner's performance compare with that of others in the group? Did the learner meet the outcomes of the clinical course and develop the essential competencies regardless of how other students in the group performed in clinical practice? Answers to these questions depend on the basis used for interpreting clinical performance, similar to interpreting test scores.

Norm-Referenced Interpretation

There are two main ways of interpreting test scores and other types of assessment results: norm referencing and criterion referencing. In norm-referenced interpretation, test scores and other assessment data are compared with those of a norm group. Norm-referenced interpretation compares a student's test scores with those of others in the class or with some other relevant group. The student's score may be described as below or above average or at a certain rank in the class. Problems with norm-referenced interpretations, for example, "grading on a curve," are that they do not indicate what the student can and cannot do, and the interpretation of a student's performance can vary widely depending on the particular comparison group selected.

In clinical settings, norm-referenced interpretations compare the student's clinical performance with the performance of a group of learners, indicating that the student has more or less clinical competence than others in the group. A clinical evaluation instrument in which student performance is rated on a scale of below to above average reflects a norm-referenced system. Again, norm-referenced clinical performance does not indicate whether a student has developed desired competencies, only whether a student performed better or worse than other students.

Criterion-Referenced Interpretation

Criterion-referenced interpretation, on the other hand, involves interpreting scores based on preset criteria, not in relation to the group of learners. With this type of measurement, an individual score is compared with a preset standard or criterion. The concern is how well the student performed and what the student can do regardless of the performance of other learners. Criterion-referenced interpretations may (a) describe the specific learning tasks a student can perform, for example, define medical terms; (b) indicate the percentage of tasks performed or items answered correctly, for example, define correctly 80% of the terms; and (c) compare performance against a set standard and decide whether the student met that standard, for example, met the medical terminology competency (Miller et al., 2013). Criterion-referenced interpretation determines how well the student

performed at the end of the instruction in comparison with the outcomes and competencies to be achieved.

With criterion-referenced clinical evaluation, student performance is compared against preset criteria. In some nursing courses, these criteria are the objectives or outcomes of the course to be met by students. In other courses, they are the competencies to be demonstrated in simulation or clinical practice, which are then used as the standards for evaluation. Rather than comparing the performance of the student with others in the group, and indicating that the student was above or below the average of the group, in criterion-referenced clinical evaluation, performance is measured against the outcomes or competencies to be demonstrated. The focus with criterion-referenced clinical evaluation is whether students achieved the outcomes of the course or demonstrated the essential clinical competencies, not how well they performed in comparison with the other students.

■ Evaluation

Evaluation is the process of making judgments about student learning and achievement, clinical performance, employee competence, and educational programs, based on the assessment data. In nursing education, evaluation typically takes the form of judging student attainment of the outcomes of the course and knowledge gained in it, and the quality of student performance in the clinical setting. With this evaluation, learning needs are identified, and additional instruction can be provided to assist students in their learning and in developing competencies for practice. Similarly, evaluation of employees provides information on their performance at varied points in time as a basis for judging their competence.

Evaluation extends beyond a test score or performance rating. Brookhart and Nitko (2019) defined *evaluation* as the process of making a value judgment about the worth or quality of a student's performance or of products developed by students representing their learning. With evaluation, the teacher makes value judgments about learners: *value* is part of the word *evaluation*. Questions, such as "How *well* did the student perform?" and "Is the student *competent* in clinical practice?" are answered by the evaluation process. The teacher collects and analyzes data about the student's performance, then makes a value judgment about the quality of that performance.

In terms of educational programs, evaluation includes collecting information *prior* to developing the program, *during* the process of program development to provide a basis for ongoing revision, and *after* implementing the program to determine its effectiveness. With program evaluation, faculty members collect data about their students, alumni, curriculum, and other dimensions of the program for the purposes of documenting the program outcomes, judging the quality of the program, and making sound decisions about curriculum revision. As educators measure outcomes for accreditation and evaluate their courses and curricula, they are engaging in program

evaluation. Although many of the concepts described in this book are applicable to program evaluation, the focus instead is on evaluating learners, including students in all types and levels of nursing programs and nurses in healthcare settings. The term *students* is used broadly to reflect both of these groups of learners.

Formative Evaluation

Evaluation fulfills two major roles: It is both formative and summative. Formative evaluation judges students' progress in meeting the desired outcomes and developing clinical competencies. With formative evaluation, the teacher judges the quality of the achievement while students are still in the process of learning (Brookhart & Nitko, 2019). Formative evaluation occurs throughout the instructional process and provides feedback for determining where further learning is needed.

With formative evaluation, the teacher assesses student learning and performance, gives students prompt and specific feedback about the knowledge and skills that still need to be acquired, and plans further instruction to enable students to fill their gaps in learning. Considering that formative evaluation is diagnostic, it typically is not graded. The purpose of formative evaluation is to determine where further learning is needed. In the classroom, formative information may be collected by teacher observation and questioning of students, diagnostic quizzes, small-group activities, written assignments, and other activities that students complete in and out of class. These same types of strategies can be used to assess student learning in online courses.

In clinical practice and other practice environments, such as simulation and skills laboratories, formative evaluation is an integral part of the instructional process. The teacher continually makes observations of students as they learn to provide patient care and develop their competencies, questions them about their understanding and decisions, discusses these observations and judgments with them, and guides them in how to improve performance. With formative evaluation, the teacher gives feedback to learners about their progress in achieving the outcomes of practice and how they can further develop their knowledge and competencies.

Summative Evaluation

Summative evaluation, on the other hand, is end-of-instruction evaluation designed to determine what the student has learned. With summative evaluation, the teacher judges the quality of the student's achievement in the course, not the progress of the learner in meeting the outcomes. Although formative evaluation occurs on a continual basis throughout the learning experience, summative evaluation is conducted on a periodic basis, for instance, every few weeks or at the midterm and final evaluation periods. This type of evaluation is "final" in nature and serves as a basis for grading and other high-stakes decisions.

Summative evaluation typically judges broader content areas and competencies than formative evaluation. Strategies used commonly for summative evaluation in the classroom and online courses are tests, papers, other assignments, and projects. In clinical practice, rating scales, written assignments, e-portfolios, projects completed about clinical experiences, and objective structured clinical examinations (OSCEs) may be used. Another strategy for summative evaluation is simulation, which can be used to assess students' decisions, skills, communication, teamwork, and other competencies.

Both formative and summative evaluation are essential components of most nursing courses. However, because formative evaluation represents feedback to learners with the goal of improving learning, it should be the major part of any nursing course. By providing feedback on a continual basis and linking that feedback with further instruction, the teacher can assist students in developing the knowledge and skills they lack.

Evaluation and Instruction

Figure 1.1 demonstrates the relationship between evaluation and instruction. The intended learning outcomes are the knowledge, skills, and competencies students are to achieve. Following assessment to determine gaps in learning and performance, the

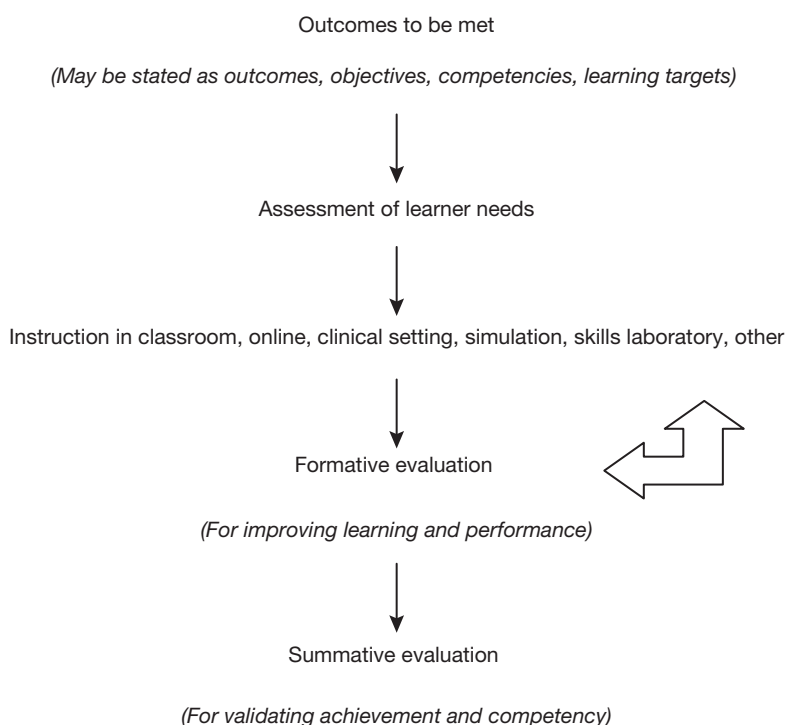


FIGURE 1.1 Relationship of evaluation and instruction.

teacher selects teaching strategies and plans clinical activities to meet those needs. This phase of the instructional process includes developing a plan for learning, selecting learning activities, and teaching learners in varied settings.

The remaining components of the instructional process relate to evaluation. Because formative evaluation focuses on judging student progress toward achieving the outcomes and demonstrating competency in clinical practice, this type of evaluation is displayed with a feedback loop to instruction. Formative evaluation provides information about further learning needs of students and where additional instruction is needed. Summative evaluation, at the end of the instruction, determines whether the outcomes have been achieved and competencies developed.

■ Outcomes for Assessment and Testing

The desired learning outcomes play an important role in teaching students in varied settings in nursing. They provide guidelines for student learning and instruction and a basis for evaluating learning. This does not mean that the teacher is unconcerned about learning that occurs but is not expressed as outcomes. Many students will acquire knowledge, skills, and values beyond those expressed in the outcomes, but the assessment strategies planned by the teacher and the evaluation that is done in a course should focus on the student learning outcomes to be achieved or competencies to be developed by students.

The knowledge, psychomotor and technical skills, and values students are to learn may be stated as *outcomes* or *competencies* to be met by students. Although there are varied definitions of *outcomes* and *competencies*, one way to think about these is that a student learning outcome is a statement that describes the knowledge, skills, or values students can demonstrate at the end of a course or another point in time (Sullivan, 2016). These are typically broad statements representing expected student learning. Competencies are more specific statements that lead to achievement of these broader learner outcomes (Scheckel, 2016). Regardless of the terms used in a particular nursing program, the outcomes and competencies provide the basis for assessing learning in the classroom, online environment, simulation and skills laboratory, and clinical setting. The next section of the chapter explains how to write outcomes as a framework for assessment.

Writing Outcomes

In earlier years, teachers developed highly specific objectives that included a description of the learner, behaviors the learner would exhibit at the end of the instruction, conditions under which the behavior would be demonstrated, and the standard of performance. An example of this format for an objective is: Given assessment data, the student identifies in writing two patient problems with supporting rationales. It is

clear from this example that highly specific instructional objectives are too prescriptive for use in nursing. Nursing students need to gain complex knowledge and skills and learn to problem solve and think critically; those outcomes cannot be identified as detailed and prescriptive objectives. In addition, specific objectives limit flexibility in planning teaching methods and in developing assessment strategies. Outcomes are less restrictive than writing specific objectives (Wittmann-Price & Fasolka, 2010). For this reason, a more general format is sufficient to express the learning outcomes and to provide a basis for assessing learning in nursing courses.

An outcome similar to the earlier objective is: The student identifies patient problems based on the assessment. This outcome, which is open-ended, provides flexibility for the teacher in developing teaching strategies and for assessing student learning. The outcome could be met and evaluated through varied activities in which students analyze assessment data, presented in a lecture, case scenario, video clip, or simulation, and then identify the patient's problems. Students might work in groups, reviewing various assessments and discussing possible problems, or they might analyze scenarios presented online. In the clinical setting, patient assignments, conferences, discussions with students, and reviews of cases provide other strategies for learners to identify patient problems from assessment data and for evaluating student competency. Stating outcomes for student learning, therefore, provides sufficient guidelines for instruction and assessment.

The outcomes are important in developing assessment strategies that collect data on the knowledge and competencies to be acquired by learners. In evaluating the sample outcome cited earlier, the method selected—for instance, a test—needs to examine student ability to identify patient problems from assessment data. The outcome does not specify the number of problems, type of problem, complexity of the assessment data, or other variables associated with the clinical situation; there is opportunity for the teacher to develop various types of test questions and assessment methods as long as they require the learner to identify patient-related problems based on the given data.

Clearly written outcomes guide the teacher in selecting assessment methods such as tests, written assignments, observations in the clinical setting, and others. If the chosen method is testing, the outcome in turn suggests the type of test item, for instance, true–false, multiple choice, or essay. In addition to guiding decisions about assessment methods, the outcome gives clues to faculty about teaching methods and learning activities to assist students in achieving the outcome. For the earlier example, teaching methods might include readings, lecture, discussion, case analysis, simulation, role-play, video clip, clinical practice, postclinical conference, and other approaches that present assessment data and ask students to identify patient problems.

Outcomes that are useful for test construction and for designing other assessment methods meet four general principles. First, the outcome should represent the

learning expected of the student at the end of the instruction. Second, it should be measurable. Terms such as *identify*, *describe*, and *analyze* are specific and may be measured; words such as *understand* and *know*, in contrast, represent a wide variety of behaviors, some simple and others complex, making these terms difficult to assess. The student's knowledge might range from identifying and naming through synthesizing and evaluating. Sample action verbs useful for writing outcomes are presented in Table 1.1.

TABLE 1.1 Sample Action Verbs for Taxonomic Levels

COGNITIVE DOMAIN	AFFECTIVE DOMAIN	PSYCHOMOTOR DOMAIN
Remembering Define Identify Label List Name Recall State	Receiving Acknowledge Ask Reply Show awareness of	Imitation Follow example of Imitate Show
Understanding Describe Differentiate Draw conclusions Explain Give examples Interpret Select Summarize	Responding Act willingly Assist Is willing to Support Respond Seek opportunities	Manipulation Assemble Carry out Follow procedure
Applying Apply Demonstrate use of Modify Predict Produce Relate Solve Use	Valuing Accept Assume responsibility Participate in Respect Support Value	Precision Demonstrate Is accurate in
Analyzing Analyze Classify Compare Contrast Differentiate Identify	Organization of values Argue Debate Declare Defend Take a stand	Articulation Adapt Carry out (accurately and in reasonable time frame) Is skillful

(continued)

TABLE 1.1 Sample Action Verbs for Taxonomic Levels (*continued*)

COGNITIVE DOMAIN	AFFECTIVE DOMAIN	PSYCHOMOTOR DOMAIN
Evaluating Appraise Assess Critique Discriminate Evaluate Judge Justify Support	Characterization by value Act consistently Stand for	Naturalization Is competent Carry out competently Integrate skill within care
Creating Construct Create Design Develop Devise Generate Plan Revise Synthesize Write		

Third, the outcomes should be as general as possible to allow for their achievement with varied course content. For instance, instead of stating that the student will identify physiological problems from the assessment of acutely ill patients, indicating that the learner will identify patient problems from assessment data provides more flexibility for the teacher in designing assessment strategies that reflect different types of problems presented in the course. Fourth, the teaching method should be omitted from the outcome to provide greater flexibility in how the instruction is planned. For example, the outcome “Uses effective communication techniques in a simulated patient–nurse interaction” limits the teacher to evaluating communication techniques through simulations rather than through interactions the student might have in the clinical setting. The outcome would be better if stated as “Uses effective communication techniques with patients.” That allows many ways of assessing whether students can communicate effectively with patients.

■ Taxonomies

The need for clearly stated outcomes and competencies, *what* the student should achieve at the end of the instruction, becomes evident when the teacher translates them into test items and other methods of assessment. Test items need to adequately assess the outcome—for instance, to identify, describe, apply, and analyze—as it

relates to the content area. Outcomes and competencies may be written to reflect three domains of learning—cognitive, affective, and psychomotor—each with its own taxonomy. The taxonomies classify the outcomes into various levels of complexity.

Cognitive Domain

The cognitive domain deals with knowledge and higher level thinking skills such as clinical reasoning and critical thinking. Learning within this domain includes the acquisition of facts and specific information, use of knowledge in practice, and higher level cognitive skills. The most widely used cognitive taxonomy was developed in 1956 by Bloom, Englehart, Furst, Hill, and Krathwohl. It includes six levels of cognitive learning, increasing in complexity: knowledge, comprehension, application, analysis, synthesis, and evaluation. This taxonomy suggests that knowledge, such as recall of specific facts, is less complex and demanding intellectually than the higher levels of learning. Evaluation, the most complex level, requires judgments based on varied criteria.

In an update of the taxonomy by Anderson and Krathwohl (2001), the names for the levels of learning were reworded as verbs, for example, the knowledge level was renamed remembering, and synthesis and evaluation were reordered. In the adapted taxonomy, the highest level of learning is creating, which is the process of synthesizing information to develop a new product (Table 1.1).

One advantage in considering this taxonomy when writing outcomes and test items is that it encourages the teacher to think about higher levels of learning expected as a result of the instruction. If the course goals reflect application of knowledge in clinical practice and complex thinking, these higher levels of learning should be reflected in the outcomes and assessment rather than focusing only on the recall of facts and other information.

In using the taxonomy, the teacher decides first on the level of cognitive learning intended and then develops outcomes and assessment methods for that particular level. Decisions about the taxonomic level at which to gear instruction and assessment depend on the teacher's judgment in considering the background of the learner, placement of the course and learning activities within the curriculum to provide for the progressive development of knowledge and competencies, and complexity of the concepts and content to be learned in relation to the time allowed for teaching. If the time for teaching and evaluation is limited, the outcomes may need to be written at a lower level. The taxonomy provides a continuum for educators to use in planning instruction and assessing learning outcomes, beginning with remembering and recalling of facts and information and progressing toward understanding, using concepts and other forms of knowledge in practice, analyzing situations, evaluating materials and situations, and creating new products.

A description and sample outcome for each of the six levels of learning in the taxonomy of the cognitive domain follow.

1. *Remembering*: Recall of facts and specific information. Memorization of specifics.
Define the term *systole*.
2. *Understanding*: Ability to describe and explain the material.
Describe the blood flow through the heart.
3. *Applying*: Use of information in a new situation. Ability to use knowledge in a new situation.
Apply evidence on nonpharmacologic interventions for managing chronic pain for patients using opioids.
4. *Analyzing*: Ability to break down material into component parts and identify the relationships among them
Analyze the unit's scheduling practices and their potential impact on patient safety and nurses' health.
5. *Evaluating*: Ability to make judgments based on criteria
Evaluate the quality and strength of evidence and applicability to practice.
6. *Creating*: Ability to develop and combine elements to create a new product
Develop guidelines for assessment, education, and follow-up care of patients with postpartum depression.

This taxonomy is useful in developing test items because it helps the teacher gear the item to a particular cognitive level. For example, if the outcome focuses on applying, the test item should measure whether the student can use the concept in a new situation, which is the intent of learning at that level. However, the taxonomy alone does not always determine the level of complexity of the item because one other consideration is how the information was presented in the instruction. For example, a test item at the application level requires use of previously learned concepts and knowledge in a new situation. Whether the situation is new for each student, however, is not known. Some students may have had clinical experience with that situation or been exposed to it through another learning activity. As another example, a question written at the understanding level may actually be at the knowledge level if the teacher used that specific explanation in class and students only need to remember the explanation to answer the item.

Affective Domain

The affective domain relates to the development of values, attitudes, and beliefs consistent with standards of professional nursing practice. Developed by Krathwohl,

Bloom, and Masia (1964), the taxonomy of the affective domain includes five levels organized around the principle of increasing involvement of the learner and internalization of a value. This principle relates to the progression of learners from mere awareness of a value, for instance, confidentiality, to internalization of that value as a basis for their own behavior. Considering the affective domain in teaching is particularly important in the clinical setting because this is where students have an opportunity to reflect on their values and use them in patient care and when collaborating with other healthcare providers. Younas and Maddigan (2019) emphasized the importance of targeting the affective domain to foster compassion among nursing students and help them to provide compassionate care.

There are two important dimensions in evaluating affective outcomes. The first relates to the student's knowledge of the values, attitudes, and beliefs that are important in guiding decisions in nursing. Prior to internalizing a value and using it as a basis for decision-making and behavior, the student needs to know what are important values in nursing. There is a cognitive base, therefore, to the development of a value system. Evaluation of this dimension focuses on acquisition of knowledge about the values, attitudes, and beliefs consistent with professional nursing practice. A variety of test items and assessment methods is appropriate to evaluate this knowledge base.

The second dimension of affective evaluation focuses on whether students have accepted these values, attitudes, and beliefs and are internalizing them for their own decision-making and behavior. Assessment at these higher levels of the affective domain is more difficult because it requires observation of learner behaviors over time to determine whether there is commitment to act according to professional values. Test items are not appropriate for these levels as the teacher is concerned with the use of values in practice and using them consistently in patient care.

A description and sample outcome for each of the five levels of learning in the affective taxonomy follow.

1. *Receiving*: Awareness of values, attitudes, and beliefs important in nursing practice. Sensitivity to a patient, clinical situation, and problem.
Express an awareness of the need for maintaining confidentiality of patient information.
2. *Responding*: Learner's reaction to a situation. Responding voluntarily to a given situation reflecting a choice made by the learner.
Share willingly feelings about caring for a dying patient.
3. *Valuing*: Internalization of a value. Acceptance of a value and the commitment to using that value as a basis for behavior.
Support the rights of patients to make their own decisions about care.

4. *Organization*: Development of a complex system of values. Creation of a value system.

Form a position about issues relating to quality and cost of care.

5. *Characterization by a value*: Internalization of a value system providing a philosophy for practice.

Act consistently to involve patients and families in care.

Psychomotor Domain

Psychomotor learning involves the development of motor skills and competency in performing clinical skills and procedures and in using technology. This domain includes activities that are movement oriented, requiring some degree of physical coordination. Motor skills have a cognitive base, which involves the principles underlying the skill. They also have an affective component reflecting the values of the nurse while carrying out the skill, for example, respecting the patient while performing the procedure.

In developing psychomotor skills, learners progress through three phases of learning: cognitive (understanding what needs to be done), associative (gradually improving performance until movements are consistent), and autonomous (performing the skill automatically; Schmidt & Lee, 2005). To progress through these levels, students need to practice the skill repetitively and receive specific, informative feedback on their performance, referred to as *deliberate practice* (Bosse et al., 2015; Ericsson, 2004; Kardong-Edgren, Oermann, & Rizzolo, 2019; McGaghie, Issenberg, Petrusa, & Scalese, 2010; Oermann, Molloy, & Vaughn, 2015). An understanding of motor skill development guides teachers in planning the instruction of skills in nursing, building in sufficient practice to gain expertise (Oermann et al., 2015; Oermann, Muckler, & Morgan, 2016).

Different taxonomies have been developed for the evaluation of psychomotor skills. One taxonomy useful in nursing education specifies five levels in the development of psychomotor skills. The lowest level is imitation learning; here the learner observes a demonstration of the skill and imitates that performance. In the second level, the learner performs the skill following written guidelines. By practicing skills, the learner refines the ability to perform them without errors (precision) and in a reasonable time frame (articulation) until they become a natural part of care (naturalization; Dave, 1970; Oermann, Shellenbarger, & Gaberson, 2018). A description of each of these levels and sample objectives follows.

1. *Imitation*: Performance of a skill following demonstration by a teacher or through multimedia. Imitative learning.

Demonstrate changing a sterile dressing.

2. *Manipulation*: Ability to follow instructions rather than needing to observe the procedure or skill.

Suction a patient according to the accepted procedure.

3. *Precision*: Ability to perform a skill accurately, independently, and without using a model or set of directions.

Take vital signs accurately.

4. *Articulation*: Coordinated performance of a skill within a reasonable time frame.

Demonstrate skill in administering an intravenous medication.

5. *Naturalization*: High degree of proficiency. Integration of skill within care.

Competently carry out skills for care of critically ill patients.

Assessment methods for psychomotor skills provide data on knowledge of the principles underlying the skill and ability to carry out the skill or procedure in simulations and with patients. Most of the evaluation of performance is done in the clinical setting, in skill and simulation laboratories, and using various technologies for distance-based clinical courses; however, test items may be used for assessing principles associated with performing the skill.

■ Use of Outcomes for Assessment and Testing

As described earlier, the taxonomies provide a framework for the teacher to plan instruction and design assessment strategies at different levels of learning, from simple to complex in the cognitive domain, from awareness of a value to developing a philosophy of practice based on a value system in the affective domain, and increasing psychomotor competency, from imitation of the skill to performance as a natural part of care. These taxonomies are of value in assessing learning and performance to gear tests and other strategies to the intended level of learning. If the outcome of learning is application, then test items also need to be at the application level. If the outcome of learning is valuing, then the assessment methods need to examine students' behaviors over time to determine whether they are committed to practice reflecting these values. If the outcome of motor skill learning is precision, then the assessment needs to focus on accuracy in performance, not the speed with which the skill is performed. The taxonomies, therefore, provide a useful framework to ensure that test items and assessment methods are at the appropriate level for the intended learning outcomes.

In developing test items and other types of assessment methods, the teacher first identifies the outcome or competency to be evaluated, then designs test items or other methods to collect information to determine whether the student has achieved it.

For the outcome “Identify characteristics of acute heart failure,” the test item would examine student ability to recall those characteristics. The expected performance is at the remembering level: recalling facts about acute heart failure, not understanding them nor using that knowledge in clinical situations.

Some teachers choose not to use outcomes as the basis for testing and evaluation and instead develop test items and other assessment methods from the content of the course. With this process, the teacher identifies explicit content areas to be evaluated; test items then sample knowledge of this content. If using this method, the teacher should refer to the course outcomes and placement of the course in the curriculum for decisions about the level of complexity of the test items and other assessment methods.

Throughout this book, multiple types of test items and other assessment methods are presented. It is assumed that these items were developed from specific outcomes or objectives, depending on the format used in the nursing program, or from explicit content areas. Regardless of whether the teacher uses outcomes or content domains as the framework for assessment, test items and other strategies should evaluate the learning outcome intended from the instruction.

■ Summary

Assessment is the collection of information for making decisions about learners, programs, and educational policies. With information collected through assessment, the teacher can determine the progress of students in a course, provide feedback to them about continued learning needs, and plan relevant instructional strategies to meet those needs and help students improve performance. Assessment provides data for making judgments about learning and performance, which is the process of evaluation, and for arriving at grades of students in courses.

A test is a set of items, each with a correct answer. Tests are a commonly used assessment strategy in nursing programs. Measurement is the process of assigning numbers to represent student achievement or performance according to certain rules, for instance, answering 20 out of 25 items correctly on a quiz. There are two main ways of interpreting assessment results: norm referencing and criterion referencing. In norm-referenced interpretation, test scores and other assessment data are interpreted by comparing them to those of other individuals. Norm-referenced clinical evaluation compares students’ clinical performance with the performance of a group of learners, indicating that the learner has more or less clinical competence than other students. Criterion-referenced interpretation, on the other hand, involves interpreting scores based on preset criteria, not in relation to a group of learners. With criterion-referenced clinical evaluation, student performance is compared with a set of criteria to be met.

Evaluation is an integral part of the instructional process in nursing. Through evaluation, the teacher makes important judgments and decisions about the extent and quality of learning. Evaluation fulfills two major roles: formative and summative. Formative evaluation judges students' progress in meeting the outcomes of learning and developing competencies for practice. It occurs throughout the instructional process and provides feedback for determining where further learning is needed. Summative evaluation, on the other hand, is end-of-instruction evaluation designed to determine what the student has learned in the classroom, an online course, or clinical practice. Summative evaluation judges the quality of the student's achievement in the course, not the progress of the learner in meeting the outcomes.

The learning outcomes and competencies play a role in teaching students in varied settings in nursing. They provide guidelines for student learning and instruction and a basis for assessing learning and performance. Evaluation serves to determine the extent and quality of the student's learning and performance in relation to these outcomes.

■ References

- Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York, NY: Longman.
- Banta, T. W., & Palomba, C. A. (2014). *Assessment essentials: Planning, implementing, and improving assessment in higher education*. San Francisco, CA: Jossey Bass, John Wiley & Sons.
- Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. White Plains, NY: Longman.
- Bosse, H. M., Mohr, J., Buss, B., Krautter, M., Weyrich, P., Herzog, W., . . . Nikendei, C. (2015). The benefits of repetitive skills training and frequency of expert feedback in the early acquisition of procedural skills. *BMC Medical Education*, 15, 22. doi:10.1186/s12909-015-0286-5
- Brookhart, S. M., & Nitko, A. J. (2019). *Educational assessment of students* (8th ed.). New York, NY: Pearson Education.
- Dave, R. H. (1970). Psychomotor levels. In R. J. Armstrong (Ed.), *Developing and writing behavioral objectives*. Tucson, AZ: Educational Innovators.
- Ericsson, K. A. (2004). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Medicine*, 79(10), S70–S81.
- Kardong-Edgren, S. K., Oermann, M. H., & Rizzolo, M. A. (2019). Emerging theories influencing the teaching of clinical nursing skills. *Journal of Continuing Education in Nursing*, 50, 257–262.
- Krathwohl, D., Bloom, B., & Masia, B. (1964). *Taxonomy of educational objectives. Handbook II: Affective domain*. New York, NY: Longman.
- McGaghie, W. C., Issenberg, S. B., Petrusa, E. R., & Scalese, R. J. (2010). A critical review of simulation-based medical education research: 2003–2009. *Medical Education*, 44, 50–63. doi:10.1111/j.1365-2923.2009.03547.x
- Miller, M. D., Linn, R. L., & Gronlund, N. E. (2013). *Measurement and assessment in teaching* (11th ed.). Upper Saddle River, NJ: Pearson Education.

- Mislevy, R. J. (2017). On measurement in educational assessment. In C. Secolsky & D. Brian Denison (Eds.), *Handbook on measurement, assessment, and evaluation in higher education* (pp. 11–31). London, England: Routledge.
- Oermann, M. H., Molloy, M., & Vaughn, J. (2015). Use of deliberate practice in teaching in nursing. *Nurse Education Today*, 35, 535–536. doi:10.1016/j.nedt.2014.11.007
- Oermann, M. H., Muckler, V. C., & Morgan, B. (2016). Framework for teaching psychomotor and procedural skills in nursing. *Journal of Continuing Education in Nursing*, 47, 278–282. doi:10.3928/00220124-20160518-10
- Oermann, M. H., Shellenbarger, T., & Gaberson, K. B. (2018). *Clinical teaching strategies in nursing* (5th ed.). New York, NY: Springer Publishing Company.
- Scheckel, M. (2016). Designing courses and learning experiences. In D. M. Billings & J. A. Halstead (Eds.), *Teaching in nursing: A guide for faculty* (5th ed., pp. 159–185). St. Louis, MO: Elsevier.
- Schmidt, R. A., & Lee, T. D. (2005). *Motor control and learning: A behavioral emphasis* (4th ed.). Champaign, IL: Human Kinetics.
- Sullivan, D. T. (2016). An introduction to curriculum development. In D. M. Billings & J. A. Halstead (Eds.), *Teaching in nursing: A guide for faculty* (5th ed., pp. 89–117). St. Louis, MO: Elsevier.
- Whittmann-Price, R. A., & Fasolka, B. J. (2010). Objectives and outcomes: The fundamental difference. *Nursing Education Perspectives*, 31, 233–236.
- Younas, A., & Maddigan, J. (2019). Proposing a policy framework for nursing education for fostering compassion in nursing students: A critical review. *Journal of Advanced Nursing* 75, 1621–1636. doi:10.1111/jan.13946

QUALITIES OF EFFECTIVE ASSESSMENT PROCEDURES: VALIDITY, RELIABILITY, AND USABILITY

How does a teacher know whether a test or another assessment instrument is good? If assessment results will be used to make important educational decisions, such as assigning grades and determining whether students are eligible for graduation, teachers must have confidence in their interpretations of test scores. Some high-stakes educational decisions have consequences for faculty members and administrators as well as students. Good assessments produce results that can be used to make appropriate inferences about learners' knowledge and abilities and thus facilitate effective decision-making. In addition, assessment tools should be practical and easy to use.

Two important questions have been posed to guide the process of constructing or proposing tests and other assessments:

1. To what extent will the interpretation of the scores be appropriate, meaningful, and useful for the intended application of the results?
2. What are the consequences of the particular uses and interpretations that are made of the results (Miller, Linn, & Gronlund, 2013, p. 70)?

This chapter explains the concept of assessment validity, the role of reliability, and their effects on the interpretive quality of assessment results. It also discusses important practical considerations that might affect the choice or development of tests and other instruments.

■ Assessment Validity

Definitions of *validity* have changed over time. Early definitions, formed in the 1940s and early 1950s, emphasized the validity of an assessment tool itself. Tests were characterized as valid or not, apart from consideration of how they were used. It was common in that era to support a claim of validity with evidence that a test correlated well with another “true” criterion. The concept of validity changed, however, in the 1950s through the 1970s to focus on evidence that an assessment tool is valid for a specific purpose. Most measurement textbooks of that era classified validity by three types—content, criterion-related, and construct—and suggested that validation of a test should include more than one approach. In the 1980s, the understanding of validity shifted again, to an emphasis on providing evidence to support the particular inferences that teachers make from assessment results. Validity was defined in terms of the appropriateness and usefulness of the inferences made from assessments, and assessment validation was seen as a process of collecting evidence to support those inferences. The usefulness of the validity “triad” also was questioned; increasingly, measurement experts recognized that construct validity was the key element and unifying concept of validity (Goodwin, 1997; Goodwin & Goodwin, 1999).

The current philosophy of validity continues to focus not on assessment tools themselves or on the appropriateness of using a test for a specific purpose, but on the meaningfulness of the interpretations that teachers make of assessment results. Tests and other assessment instruments yield scores that teachers use to make inferences about how much learners know or what they can do. Validity refers to the adequacy and appropriateness of those interpretations and inferences and how the assessment results are used (Miller et al., 2013). The emphasis is on the consequences of measurement: Does the teacher make accurate interpretations about learners’ knowledge or ability based on their assessment scores? Assessment experts increasingly suggest that in addition to collecting evidence to support the accuracy of inferences made, evidence also should be collected about the intended and unintended consequences of the use of a test (Brookhart & Nitko, 2019; Goodwin, 1997; Goodwin & Goodwin, 1999).

Validity does not exist on an all-or-none basis (Miller et al., 2013); there are degrees of validity depending on the purpose of the assessment and how the results are to be used. A given assessment may be used for many different purposes, and inferences about the results may have greater validity for one purpose than for another. For example, a test designed to measure knowledge of perioperative nursing guidelines may produce results that have high validity for the purpose of determining certification for perioperative staff nurses, but the results may have low validity for assigning grades to students in a perioperative nursing elective course. In addition, validity evidence may change over time, so validation of inferences must not be considered a one-time event.

No one assessment will produce results that are perfectly valid for a given purpose. Combining results from several different types of assessments, such as tests, written assignments, and class participation, improves the validity of the decisions made about students' attainments. In addition, weighing one assessment outcome too heavily in relation to others, such as basing course grades almost exclusively on test scores, results in lowered validity (Brookhart & Nitko, 2019).

Validity now is considered a unitary concept (Brookhart & Nitko, 2019; Miller et al., 2013). The concept of validity in testing is described in the *Standards for Educational and Psychological Testing* prepared by a joint committee of the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). The most recent *Standards* (2014) no longer includes the view that there are different types of validity—for example, construct, criterion-related, and content.

Instead, there is a variety of sources of evidence to support the validity of the interpretation and use of assessment results. The strongest case for validity can be made when evidence is collected regarding four major considerations for validation:

1. Content
2. Construct
3. Assessment–criterion relationships
4. Consequences (Miller et al., 2013, p. 74)

Each of these considerations is discussed as to how it can be used in nursing education settings.

Content Considerations

The goal of content validation is to determine the degree to which a sample of assessment tasks accurately represents the domain of content or abilities about which the teacher wants to interpret assessment results. Tests and other assessment measures usually contain only a sample of all possible items or tasks that could be used to assess the domain of interest. However, interpretations of assessment results are based on what the teacher believes to be the universe of items that could have been generated. In other words, when a student correctly answers 83% of the items on a women's health nursing final examination, the teacher usually infers that the student probably would answer correctly 83% of all items in the universe of women's health nursing content. The test score thus serves as an indicator of the student's true standing in the larger domain. Although this type of generalization is commonly made, it should be noted that the domains of achievement in nursing education involve complex understandings and integrated performances, about which it is difficult to judge the representativeness of a sample of assessment tasks (Miller et al., 2013).

A superficial conclusion could be made about the match between a test's appearance and its intended use by asking a panel of experts to judge whether the test appears to be based on appropriate content. This type of judgment, sometimes referred to as *face validity*, is not sufficient evidence of content representativeness and should not be used as a substitute for rigorous appraisal of sampling adequacy (Miller et al., 2013).

Efforts to include suitable content on an assessment can and should be made during its development. This process begins with defining the universe of content. The content definition should be related to the purpose for which the test will be used. For example, if a test is supposed to measure a new staff nurse's understanding of hospital safety policies and procedures presented during orientation, the teacher first defines the universe of content by outlining the knowledge about policies that the staff nurse needs to function satisfactorily. The teacher then uses professional judgment to write or select test items that satisfactorily represent this desired content domain. A system for documenting this process, the construction of a test blueprint or table of specifications, is described in Chapter 3, Planning for Testing.

If the teacher needs to select an appropriate assessment for a particular use, for example, choosing a standardized achievement test, content validation is also of concern. A published test may or may not be suitable for the intended use in a particular nursing education program or with a specific group of learners. The ultimate responsibility for appropriate use of an assessment and interpretation of results lies with the teacher (AERA, APA, & NCME, 2014; Miller et al., 2013). To determine the extent to which an existing test is suitable, experts in the domain review the assessment, item by item, to determine whether the items or tasks are relevant and satisfactorily represent the defined domain, represented by the table of specifications, and the desired learning outcomes. Because these judgments admittedly are subjective, the trustworthiness of this evidence depends on clear instructions to the experts and estimation of rater reliability.

Construct Considerations

Construct validity has been proposed as the “umbrella” under which all types of assessment validation belong (Goodwin, 1997; Goodwin & Goodwin, 1999). Content validation determines how well test scores represent a given domain and is important in evaluating assessments of achievement. When teachers need to make inferences from assessment results to more general abilities and characteristics, however, such as clinical reasoning or communication ability, a critical consideration is the construct that the assessment is intended to measure (Miller et al., 2013).

A *construct* is an individual characteristic that is assumed to exist because it explains some observed behavior. As a theoretical construction, it cannot be observed directly, but it can be inferred from performance on an assessment. *Construct validation* is the

process of determining the extent to which assessment results can be interpreted in terms of a given construct or set of constructs. Two questions, applicable to both teacher-constructed and published assessments, are central to the process of construct validation:

1. How adequately does the assessment represent the construct of interest (construct representation)?
2. Is the observed performance influenced by any irrelevant or ancillary factors (construct relevance)? (Miller et al., 2013, p. 81)

Assessment validity is reduced to the extent that important elements of the construct are underrepresented in the assessment. For example, if the construct of interest is clinical problem-solving ability, the validity of a clinical performance assessment would be weakened if it focused entirely on problems defined by the teacher, because the learner's ability to recognize and define clinical problems is an important aspect of clinical problem-solving (Gaberson & Oermann, 2018).

The influence of factors that are unrelated or irrelevant to the construct of interest also reduces assessment validity (Brookhart & Nitko, 2019). For example, students who are non-native English speakers may perform poorly on an assessment of clinical problem-solving, not because of limited ability to recognize, identify, and solve problems, but because of unfamiliarity with language or cultural colloquialisms used by patients or teachers (Bosher, 2009; Bosher & Bowles, 2008). Another potential construct-irrelevant factor is writing skill. For example, the ability to communicate clearly and accurately in writing may be an important outcome of a nursing education program, but the construct of interest for a course writing assignment is clinical problem-solving. To the extent that student scores on that assignment are affected by spelling or grammatical errors, the construct-relevant validity of the assessment is reduced. Testwiseness, performance anxiety, and learner motivation are additional examples of possible construct-irrelevant factors that may undermine assessment validity (Miller et al., 2013).

Construct validation for a teacher-made assessment occurs primarily during its development by collecting evidence of construct representation and construct relevance from a variety of sources. Test manuals for published tests should include evidence that these methods were used to generate evidence of construct validity. Methods used in construct validation include:

1. *Defining the domain to be measured.* The assessment specifications should clearly define the meaning of the construct so that it is possible to judge whether the assessment includes relevant and representative tasks.
2. *Analyzing the process of responding to tasks required by the assessment.* The teacher can administer an assessment task to the learners (e.g., a multiple-choice item that purportedly assesses clinical reasoning) and ask them

to think aloud while they perform the test (e.g., explain how they arrived at the answer they chose). This method may reveal that students were able to identify the correct answer because the same example was used in class or in an assigned reading, not because they were able to analyze the situation critically.

3. *Comparing assessment results of known groups.* Sometimes it is reasonable to expect that scores on a particular measure will differ from one group to another because members of those groups are known to possess different levels of the ability being measured. For example, if the purpose of a test is to measure students' ability to solve pediatric clinical problems, students who achieve high scores on this test would be assumed to be better problem solvers than students who achieve low scores. To collect evidence in support of this assumption, the teacher might design a study to determine whether student scores on the test are correlated with their scores on a standardized test of clinical problem-solving in nursing. The teacher could divide the sample of students into two groups based on their standardized test scores: those who scored high on the standardized test in one group and those whose standardized test scores were low in the other group. Then the teacher would compare the teacher-made test scores of the students in both groups. If the teacher's hypothesis is confirmed (i.e., if the students with high-standardized test scores obtained high scores on the teacher-made test), this evidence could be used as partial support for construct validation (Miller et al., 2013).
Group-comparison techniques also have been used in studies of test bias or test fairness. Approaches to detection of test bias have looked for differential item functioning (DIF) related to test-takers' race, gender, or culture. If test items function differently for members of groups with characteristics that do not directly relate to the variable of interest, differential validity of inferences from the test scores may result. Issues related to test bias are discussed more fully in Chapter 16, Social, Ethical, and Legal Issues.
4. *Comparing assessment results before and after a learning activity.* It is reasonable to expect that assessments of student performance would improve during instruction, whether in the classroom or in the clinical area, but assessment results should not be affected by other variables such as anxiety or memory of the preinstruction assessment content. For example, evidence that assessment scores improve following instruction but are unaffected by an intervention designed to reduce students' test anxiety would support the assessment's construct validity (Miller et al., 2013).
5. *Correlating assessment results with other measures.* Scores produced by a particular assessment should correlate well with scores of other measures of the

same construct but show poor correlation with measures of a different construct. For example, teachers' ratings of students' performance in pediatric clinical settings should correlate highly with scores on a final exam testing knowledge of nursing care of children, but may not correlate satisfactorily with their classroom or clinical performance in a women's health course. These correlations may be used to support the claim that a test measures the construct of interest (Miller et al., 2013).

Assessment–Criterion Relationship Considerations

This approach to obtaining validity evidence focuses on predicting future performance (the criterion) based on current assessment results. For example, nursing faculties often use scores from a standardized comprehensive exam given in the final academic semester or quarter to predict whether prelicensure students are likely to be successful on the NCLEX® (National Council Licensure Examination; the criterion measure). Obtaining this type of evidence involves a *predictive* validation study (Miller et al., 2013).

If teachers want to use assessment results to estimate students' performance on another assessment (the criterion measure) at the same time, the validity evidence is *concurrent*, and obtaining this type of evidence requires a concurrent validation study. This type of evidence may be desirable for making a decision about whether one test or measurement instrument may be substituted for another, more resource-intensive one. For example, a staff development educator may want to collect concurrent validity evidence to determine whether a checklist with a rating scale can be substituted for a less efficient narrative appraisal of a staff nurse's competence.

Teachers rarely conduct formal studies of the extent to which the scores on assessments that they have constructed are correlated with criterion measures. In some cases, adequate criterion measures are not available; the test in use is considered to be the best instrument that has been devised to measure the ability in question. If better measures were available, they might be used instead of the test being validated. However, for tests with high-stakes outcomes, such as licensure and certification, this type of validity evidence is crucial. Multiple criterion measures often are used so that the strengths of one measure may offset the weaknesses of others.

The relationship between assessment scores and those obtained on the criterion measure usually is expressed as a correlation coefficient. A desired level of correlation between the two measures cannot be recommended because the correlation may be influenced by a number of factors, including test length, variability of scores in the distribution, and the amount of time between measures. The teacher who uses the test must use good professional judgment to determine what magnitude of correlation is considered adequate for the intended use of the assessment for which criterion-related evidence is desired.

Consideration of Consequences

Incorporating concern about the consequences of assessment into the concept of validity is a relatively recent trend. Assessment has both intended and unintended consequences, and teachers and administrators must consider those consequences when judging whether or not they are using the results validly (Brookhart & Nitko, 2019). For example, the faculties of many undergraduate nursing programs have adopted programs of achievement testing that are designed to assess student performance throughout the nursing curriculum. The intended positive consequence of such testing is to identify students at risk of failure on the NCLEX, and to use this information to design remediation programs to increase student learning. Unintended negative consequences, however, may include increased student anxiety, decreased time for instruction relative to increased time allotted for testing, and tailoring instruction to more closely match the content of the tests while focusing less intently on other important aspects of the curriculum that will not be tested on the NCLEX. The intended consequence of using standardized comprehensive exam scores to predict success on the NCLEX may be to motivate students whose assessment results predict failure to remediate and prepare more thoroughly for the licensure exam. But an unintended consequence might be that students whose comprehensive exam scores predict NCLEX success may decide not to prepare further for that important exam, risking a negative outcome.

Ultimately, assessment validity requires an evaluation of interpretations and use of assessment results. The concept of validity thus has expanded to include consideration of the consequences of assessment use and how results are interpreted to students, teachers, and other stakeholders. An adequate consideration of consequences must include both intended and unintended effects of assessment, particularly when assessment results are used to make high-stakes decisions (Miller et al., 2013).

Influences on Validity

A number of factors affect the validity of assessment results, including characteristics of the assessment itself, the administration and scoring procedures, and the test-takers. Teachers should be alert to these factors when constructing assessments or choosing published ones (Miller et al., 2013).

Characteristics of the Assessment

Many factors can prevent the assessment items or tasks from functioning as intended, thereby decreasing the validity of the interpretations of the assessment results. Such factors include unclear directions, ambiguous statements, oversampling of easy-to-assess aspects, too few assessment items, poor arrangement of assessment items, an

obvious pattern of correct answers, and clerical errors in test construction (Miller et al., 2013). Ways to prevent test-construction errors such as these are addressed in Chapters 3, 4, 5, 6, 7, and 10.

Assessment Administration and Scoring Factors

On teacher-made assessments, factors such as insufficient time, inconsistency in giving aid to students who ask questions during the assessment, cheating, and scoring errors may lower validity. On published assessments, an additional factor may be failure to follow the standard directions, including time limits (Miller et al., 2013).

Student Characteristics

Some invalid interpretations of assessment results are the result of personal factors that influence a student's performance on the assessment. For example, a student may have had an emotionally upsetting event such as an auto accident or death in the family just prior to the assessment, test anxiety may prevent the student from performing according to true ability level, or the student may not be motivated to exert maximum effort on the assessment. These and similar factors may modify student responses on the assessment and distort the results, leading to lower validity (Miller et al., 2013).

■ Reliability

Reliability refers to the consistency of assessment results. If an assessment produces reliable scores, the same group of students would achieve approximately the same scores if the same assessment were given on another occasion, assuming that no further learning had taken place during the time interval. Each assessment produces a limited measure of performance at a specific time. If this measurement is reasonably consistent over time, with different raters, or with different samples of the same domain, teachers can be more confident in the assessment results.

Perfect consistency is indicated by a reliability coefficient of 1.00. However, this value is virtually never obtained in real educational settings. Standardized achievement tests usually have reliability coefficients in the .85 to .95 range (Brookhart & Nitko, 2019), but teacher-made tests rarely demonstrate this level of consistency because many extraneous factors may influence the measurement of performance.

Assessment results may be inconsistent because:

1. The behavior being measured is unstable over time because of fluctuations in memory, attention, and effort; intervening learning experiences; or varying emotional or health status.

2. The sample of tasks varies from one assessment to another, and some students find one assessment to be easier than another because it contains tasks related to topics they know well.
3. Assessment conditions vary significantly between assessments.
4. Scoring procedures are inconsistent (the same rater may use different criteria on different assessments, or different raters may not reach perfect agreement on the same assessment).

These and other factors introduce a certain but unknown amount of error into every measurement. Methods of determining assessment reliability, therefore, are means of estimating how much measurement error is present under varying assessment conditions. When assessment results are reasonably consistent, there is less measurement error and greater reliability (Miller et al., 2013).

For purposes of understanding sources of inconsistency, it is helpful to view an assessment score as having two components, a true score and an error score, represented by the following equation:

$$X = T + E \quad (2.1)$$

A student's actual assessment score (X) is also known as the observed or obtained score. That student's hypothetical true score (T) cannot be measured directly because it is the average of all scores the student would obtain if tested on many occasions with the same test. The observed score contains a certain amount of measurement error (E), which may be a positive or a negative value. This error of measurement, representing the difference between the observed score and the true score, results in a student's obtained score being higher or lower than his or her true score (Brookhart & Nitko, 2019). If it were possible to measure directly the amount of measurement error that occurred on each testing occasion, two of the values in this equation would be known (X and E), and we would be able to calculate the true score (T). However, we can only estimate indirectly the amount of measurement error, leaving us with a hypothetical true score. Therefore, teachers need to recognize that the obtained score on any test is only an estimate of what the student really knows about the domain being tested.

For example, Matt may obtain a higher score than Kelly on a community health nursing unit test because Matt truly knows more about the content than Kelly does. Test scores should reflect this kind of difference, and if the difference in knowledge is the only explanation for the score difference, no error is involved. However, there may be other potential explanations for the difference between Kelly's and Matt's test scores. Matt may have behaved dishonestly to obtain a copy of the test in advance; knowing which items would be included, he had the opportunity to use this unauthorized resource to determine the correct answers to those items. In his case, measurement error would have *increased* Matt's obtained score. Kelly may have worked

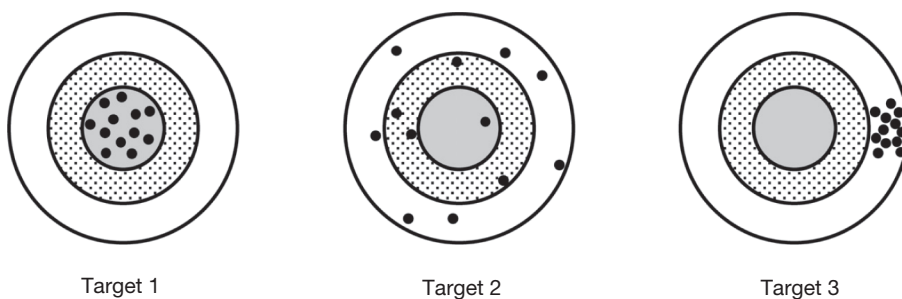
overtime the night before the test and may not have gotten enough sleep to allow her to feel alert during the test. Thus, her performance may have been affected by her fatigue and her decreased ability to concentrate, resulting in an obtained score *lower* than her true score. One goal of assessment designers, therefore, is to maximize the amount of score variance that explains real differences in ability and to minimize the amount of random error variance of scores.

The following points further explain the concept of assessment reliability (Brookhart & Nitko, 2019; Miller et al., 2013):

1. *Reliability pertains to assessment results, not to the assessment instrument itself.* The reliability of results produced by a given instrument will vary depending on the characteristics of the students being assessed and the circumstances under which it is used. Reliability should be estimated with each use of an assessment instrument.
2. *A reliability estimate always refers to a particular type of consistency.* Assessment results may be consistent over different periods of time, or different samples of the domain, or different raters or observers. It is possible for assessment results to be reliable in one or more of these respects but not in others. The desired type of reliability evidence depends on the intended use of the assessment results. For example, if the faculty wants to assess students' ability to make sound clinical decisions in a variety of settings, a measure of consistency over time would not be appropriate. Instead, an estimate of consistency of performance across different tasks would be more useful.
3. *A reliability estimate always is calculated with statistical indices.* Consistency of assessment scores over time, among raters, or across different assessment measures involves determining the relationship between two or more sets of scores. The extent of consistency is expressed in terms of a reliability coefficient (a form of correlation coefficient) or a standard error of measurement (*SEM*). A reliability coefficient differs from a validity coefficient (described earlier) in that it is based on agreement between two sets of assessment results from the same procedure instead of agreement with an external criterion.
4. *Reliability is an essential but insufficient condition for validity.* Teachers cannot make valid inferences from inconsistent assessment results. Conversely, highly consistent results may indicate only that the assessment measured the wrong construct (although doing it very reliably). Thus, low reliability always produces a low degree of validity, but a high reliability estimate does not guarantee a high degree of validity. "In short, reliability merely provides the consistency that makes validity possible" (Miller et al., 2013, p. 110).

An example may help to illustrate the relationship between validity and reliability. Suppose that the author of this chapter was given a test of her knowledge of assessment principles. The author of a textbook on assessment in nursing education might be expected to achieve a high score on such a test. However, if the test were written in Mandarin Chinese, the author's score probably would be very low, even if she were a remarkably good guesser, because she cannot read Mandarin Chinese. If the same test were administered the following week, and every week for a month, her scores would likely be consistently low, assuming that she had not learned Mandarin Chinese in the intervals between tests. Therefore, these test scores would be considered reliable because there would be a high correlation among scores obtained on the same test over a period of several administrations. But a valid score-based interpretation of the author's knowledge of assessment principles could not be drawn because the test was not appropriate for its intended use.

Figure 2.1 uses a target-shooting analogy to further illustrate these relationships. When they design and administer assessments, teachers attempt to consistently (reliably) measure the true value of what students know and can do (hit the bull's-eye); if they succeed, they can make valid inferences from assessment results. Target 1 illustrates the reliability of scores that are closely grouped on the bull's-eye, the true score, allowing the teacher to make valid inferences about them. Target 2 displays assessment scores that are widely scattered at a distance from the true score; these scores are not reliable, contributing to a lack of validity evidence. Target 3 shows assessment scores that are reliable because they are closely grouped together, but they are still distant from the true score. The teacher would not be able to make valid interpretations of such scores (Miller et al., 2013).



Reliability (consistency) is needed to obtain valid results
(but one can be consistently “off target”).

FIGURE 2.1 The relationship between reliability and validity.

Source: From Miller, M. D., Linn, R. L., & Gronlund, N. E. (2013). *Measurement and assessment in teaching* (11th ed.). Upper Saddle River, NJ: Pearson Education. Copyright © 2013 by Pearson Education. Reprinted by permission of the publisher.

Methods of Estimating Reliability

Because reliability is viewed in terms of different types of consistency, these types are determined by different methods: over time (stability), among different forms of the assessment (equivalence), within the assessment itself (internal consistency), and among different raters (consistency of ratings or interrater reliability). Each method of estimating reliability is described in further detail.

Measure of Stability

Evidence of stability indicates whether students would achieve essentially the same scores if they took the same assessment at another time—a test–retest procedure. The correlation between the set of scores obtained on the first administration and the set obtained on the second yields a test–retest reliability coefficient. This type of reliability evidence is known as *stability*, and is appropriate for situations in which the trait being measured is expected to be stable over time. In general, the longer the period of time between administrations of the test, the lower the stability–reliability estimate (Brookhart & Nitko, 2019). In nursing education settings, the test–retest method of obtaining reliability information may have limited usefulness. If the same test items are used on both tests, the students’ answers on the retest are not independent of their answers on the first test. That is, their responses to the second test may be influenced to some extent by recall of their previous responses or by discussion or individual review of content after taking the first test. In addition, if there is a long interval between testing occasions, other factors such as real changes in student ability as a result of learning may affect the retest scores. When selecting standardized tests, however, stability is an important consideration (Miller et al., 2013). Test–retest reliability coefficients for standardized achievement tests should be very high (Brookhart & Nitko, 2019).

Measure of Equivalence

Equivalent-forms reliability, also known as *alternate* or *parallel forms*, involves the use of two or more forms of the same assessment, constructed independently but based on the same set of specifications. Both forms of the assessment are administered to the same group of students in close succession, and the resulting scores are correlated. A high reliability coefficient indicates that the two forms sample the domain of interest equally well, and that generalizations about student performance from one assessment to the other can be made with a high degree of validity. The equivalent-form estimates of reliability are widely used in standardized testing, primarily to ensure test security, but the user cannot assume comparability of alternate forms unless the test manual provides information about equivalence (Miller et al., 2013). This method of reliability estimation is not practical for

teacher-constructed assessments because most teachers do not find time to prepare two forms of the same test, let alone to ensure that these forms indeed are equivalent (Brookhart & Nitko, 2019).

Measures of Internal Consistency

Internal consistency methods can be used with a set of scores from only one administration of a single assessment. Sometimes referred to as *split-half* or *half-length* methods, estimates of internal consistency reveal the extent to which consistent results are obtained from two halves of the same assessment.

The split-half technique consists of dividing the assessment into two equal subtests, usually by including odd-numbered items on one subtest and even-numbered items on the other. Then the subtests are scored separately, and the two subscores are correlated. The resulting correlation coefficient is an estimate of the extent to which the two halves consistently perform the same measurement. Longer assessments tend to produce more reliable results than shorter ones, in part because they tend to sample the content domain more fully. Therefore, a split-half reliability estimate tends to underestimate the true reliability of the scores produced by the whole assessment (because each subset includes only half of the total number of items). This underestimate can be corrected by using the Spearman–Brown prophecy formula, also called the Spearman–Brown double length formula, as represented by the following equation (Miller et al., 2013, p. 115):

$$\text{Reliability of full assessment} = \frac{2 \times \text{Correlation between half test scores}}{1 + \text{Correlation between half test scores}} \quad (2.2)$$

Another method of estimating the internal consistency of a test is to use certain types of coefficient alpha. Coefficient alpha reliability estimates provide information about the extent to which the assessment tasks measure similar characteristics. When the assessment contains relatively homogenous material, the coefficient alpha reliability estimate is similar to that produced by the split-half method. In other words, coefficient alpha represents the average correlation obtained from all possible split-half reliability estimates. The Kuder–Richardson formulas are a specific type of coefficient alpha. Computation of Formula 20 (K-R20) is based on the proportion of correct responses and the standard deviation of the total score distribution. If the assessment items are not expected to vary much in difficulty, the simpler Formula 21 (K-R21) can be used to approximate the value of K-R20, although in most cases it will produce a slightly lower estimate of reliability. To use either formula, the assessment items must be scored dichotomously, that is, right or wrong (Brookhart & Nitko, 2019; Miller et al., 2013). If the assessment items could receive a range of points, coefficient alpha should be used to provide a reliability estimate. The widespread

availability of computer software for assessment scoring and test and item analysis makes these otherwise cumbersome calculations more feasible to obtain efficiently (Miller et al., 2013).

Measures of Consistency of Ratings

Depending on the type of assessment, error may arise from the procedures used to score a test. Teachers may need to collect evidence to answer the question, “Would this student have obtained the same score if a different person had scored the assessment or judged the performance?” The easiest method for collecting this evidence is to have two equally qualified persons score each student’s paper or rate each student’s performance. The two scores then are compared to produce a percentage of agreement or correlated to produce an index of scorer consistency, depending on whether agreement in an absolute (actual score level) sense or a relative (rank order) sense is required (Brookhart & Nitko, 2019). In nursing education programs, students commonly are evaluated on the basis of their actual score level; this is the usual basis for assigning grades and determining whether students have satisfied requirements for progressing through the program and meeting graduation requirements. Judging performance in a relative sense would only rank students from highest to lowest assessment results. Agreement on rank would not necessarily provide the appropriate evidence for determining whether students have met intended learning outcomes.

Achieving a high degree of interrater consistency depends on consensus of judgment among raters regarding the value of a given performance. Such consensus is facilitated by the use of scoring rubrics and training of raters to use those rubrics. Interrater consistency is important to ensure that differences in stringency or leniency of ratings between raters do not place some students at a disadvantage (Miller et al., 2013).

Standard Error of Measurement

To describe the inconsistency of assessment scores, we could assess students repeatedly and note how much the scores vary. (We could only do this hypothetically, however, because students’ abilities might change during the process of repeated assessments.) Repeated assessments would yield a distribution of each student’s obtained scores that cluster around a mean value. This mean is the student’s true score. The standard deviation of this distribution is the standard error of measurement (*SEM*).

Because we cannot repeatedly reassess students without changing the trait we are assessing, we estimate the *SEM* using the following equation:

$$SEM = SD\chi\sqrt{1 - \text{Reliability coefficient}} \quad (2.3)$$

where $SD\chi$ is the standard deviation of the obtained scores in a distribution. Therefore, the numerical value of *SEM* estimates the amount by which a student’s

observed score is likely to deviate from the true score. For example, if the $SEM = 4.0$, a student's obtained score is likely to be about 4 points above or below the true score. Because of this likely deviation from the true score, we must interpret the student's obtained score on the assessment procedure as only an estimate of that student's true score (Brookhart & Nitko, 2019).

The size of the SEM depends on both the reliability estimate and the standard deviation of the obtained scores. When the reliability coefficient is held constant, the SEM becomes larger as the SD_{χ} increases and smaller as the SD_{χ} decreases. When the SD_{χ} is held constant, the SEM becomes smaller as the reliability coefficient increases and larger as the reliability coefficient decreases. Smaller $SEMs$ indicate that the observed assessment scores are very near a student's true score. Keep in mind that the type of reliability coefficient used in the SEM formula estimates the reliability for a specific type of measurement error (Brookhart & Nitko, 2019).

Factors That Influence the Reliability of Scores

From the previous discussion, it is obvious that various factors can influence the reliability of a set of test scores. These factors can be categorized into three main sources: the assessment instrument itself, the student, and the assessment administration conditions.

Assessment-related factors include the length of the test, the homogeneity of assessment tasks, and the difficulty and discrimination ability of the individual items. In general, the greater the number of assessment tasks (e.g., test items), the greater the score reliability. The Spearman–Brown reliability estimate formula can be used to estimate the effect on the reliability coefficient of adding assessment tasks. For example, if a 10-item test has a reliability coefficient of .40, adding 15 items (creating a test that is 2.5 times the length of the original test) would produce a reliability estimate of .625. Of course, adding assessment tasks to increase score reliability may be counterproductive after a certain point. After that point, adding tasks will increase the reliability only slightly, and student fatigue and boredom actually may introduce more measurement error. Score reliability also is enhanced by homogeneity of content covered by the assessment. Course content that is tightly organized and highly interrelated tends to make homogeneous assessment content easier to achieve. Finally, the technical quality of assessment items, their difficulty, and their ability to discriminate between students who know the content and students who don't also affects the reliability of scores. Moderately difficult items that discriminate well between high achievers and low achievers and that contain no technical errors contribute a great deal to score reliability. See Chapter 12, *Test and Item Analysis: Interpreting Test Results*, for a discussion of item difficulty and discrimination.

Student-related factors include the heterogeneity of the student group, test-taking ability, and motivation. In general, reliability tends to increase as the range of talent

in the group of students increases. Therefore, in situations in which students are very similar to one another in ability, such as in graduate programs, assessments are likely to produce scores with somewhat lower reliability than desired. A student's test-taking skill and experience also may influence score reliability to the extent that the student is able to obtain a higher score than true ability would predict. The effect of motivation on reliability is proportional to the extent to which it influences individual students differently. If some students are not motivated to put forth their best efforts on an assessment, their actual achievement levels may not be accurately represented, and their relative achievement in comparison to other students will be difficult to judge.

Teachers need to control assessment administration conditions to enhance the reliability of scores. Inadequate time to complete the assessment can lower the reliability of scores because some students who know the content well will be unable to respond to all of the items. Cheating also contributes random errors to assessment scores when students are able to respond correctly to items to which they actually do not know the answer. Cheating, therefore, has the effect of raising the offenders' observed scores above their true scores, contributing to inaccurate and less meaningful interpretations of test scores.

Because a reliability coefficient is an indication of the amount of measurement error associated with a set of scores, it is useful information for evaluating the meaning and usefulness of those scores. Again, it is important to remember that the numerical value of a reliability coefficient is not a stable property of an assessment; it will fluctuate from one sample of students to another each time the assessment is administered. Teachers often wonder how high the reliability coefficient should be to ensure that an assessment will produce reliable results. The degree of reliability desired depends on a number of factors, including the importance of the educational decision being made, how far-reaching the consequences would be, and whether it is possible to confirm or reverse the judgment later. The more important and the less reversible the decision is based on the assessment results, the higher the reliability should be (Brookhart & Nitko, 2019). For irreversible decisions that would have serious consequences, like the results of the first attempt of the NCLEX, a high degree of reliability must be assured. For less important decisions, especially if later review can confirm or reverse them without serious harm to the student, less reliable methods may be acceptable. For teacher-made assessments, a reliability coefficient between .60 and .85 is desirable (Miller et al., 2013).

■ Practicality

Although reliability and validity are used to describe the ways in which scores are interpreted and used, *practicality* (also referred to as *usability*) is a quality of the assessment instrument itself and its administration procedures. Assessment

procedures should be efficient and economical. An assessment is practical or usable to the extent that it is easy to administer and score, does not take too much time away from other instructional activities, and has reasonable resource requirements. Whether they develop their own tests and other measurement tools or use published instruments, teachers should focus on the following questions to help guide the selection of appropriate assessment procedures (Brookhart & Nitko, 2019; Miller et al., 2013):

1. *Is the assessment easy to construct and use?* Essay test items may be written more quickly and easily than multiple-choice items, but they will take more time to score. Multiple-choice items that assess a student's ability to think critically about clinical problems are time-consuming to construct, but they may be machine-scored quickly and accurately. The teacher must determine the best use of the time available for assessment construction, administration, and scoring. If a published test is selected for assessment of students' competencies just prior to graduation, is it practical to use? Does proper administration of the test require special training? Are the test administration directions easy to understand?
2. *Is the time needed to administer and score the assessment and interpret the results reasonable?* A teacher of a 15-week, 3-credit course wants to give a weekly 10-point quiz that would be reviewed immediately and self-scored by students; these procedures would take a total of 30 minutes of class time. Is this the best use of instructional time? The teacher may decide that there is enormous value in the immediate feedback provided to students during the test review, and that the opportunity to obtain weekly information about the effectiveness of instruction is also beneficial; to that teacher, 30 minutes weekly is time well spent on assessment. Another teacher, whose total instructional time is only 4 days, may find that administering more than one test consumes time that is needed for teaching. Evaluation is an important step in the instructional process, but it cannot replace teaching. Although students often learn from the process of preparing for and taking assessments, instruction is not the primary purpose of assessment, and assessment is not the most efficient or effective way to achieve instructional goals. On the other hand, reliability is related to the length of an assessment (i.e., the number of assessment tasks); it may be preferable to use fewer assessments of longer length rather than more frequent shorter assessments.
3. *Are the costs associated with assessment construction, administration, and scoring reasonable?* Although teacher-made assessments may seem to be less expensive than published instruments, the cost of the instructor's time spent in assessment development must be taken into consideration. Additional costs associated with the scoring of teacher-made assessments also must be

calculated. What is the initial cost of purchasing test booklets for published instruments, and can test booklets be reused? What is the cost of answer sheets, and does that cost include scoring services? When considering the adoption of a computerized testing package, teachers and administrators must decide how the costs of the program will be paid and by whom (the educational program or the individual students).

4. *Can the assessment results be interpreted easily and accurately by those who will use them?* If teachers score their own assessments, will they obtain information that will help them to interpret the results accurately? For example, will they have test and item statistics that will help them make meaning out of the individual test scores? Scanners and software are available that will quickly score assessments that use certain types of answer sheets, but the scope of the information produced in score reports varies considerably. Purchased assessments that are scored by the publisher also yield reports of test results. Are these reports useful for their intended purpose? What information is needed or desired by the teachers who will make evaluation decisions, and is that information provided by the score-reporting service?

Examples of information on score reports include individual raw total scores, individual raw subtest scores, group mean and median scores, individual or group profiles, and individual standard scores. Will the teachers who receive the reports need special training to interpret this information accurately? Some assessment publishers restrict the purchase of instruments to users with certain educational and experience qualifications, in part so that the test results will be interpreted and used properly.

■ Summary

Because assessment results often are used to make important educational decisions, teachers must have confidence in their interpretations of test scores. Assessment validity produces results that permit teachers to make accurate interpretations about a test-taker's knowledge or ability. Validity is not a static property of the assessment itself; rather, it refers to the ways in which teachers interpret and use the assessment results. Validity is not an either/or judgment; there are degrees of validity depending on the purpose of the assessment and how the results are to be used. A single assessment may be used for many different purposes, and the results may have greater validity for one purpose than for another.

Teachers must gather a variety of sources of evidence to support the validity of their interpretation and use of assessment results. Four major considerations for validation are related to content, construct, assessment–criterion relationships, and the consequences of assessment. *Content considerations* focus on the extent to which the sample of assessment items or tasks represents the domain of content or abilities

that the teacher wants to measure. Content validity evidence may be obtained during the assessment–development process as well as by appraising a completed assessment, as in the case of a purchased instrument. Currently, *construct considerations* are seen as the unifying concept of assessment validity, representing the extent to which score-based inferences about the construct of interest are accurate and meaningful. Two questions central to the process of construct validation concern how adequately the assessment represents the construct of interest (construct representation), and the extent to which irrelevant or ancillary factors influence the results (construct relevance). Methods used in construct validation include defining the domain to be measured, analyzing the task–response processes required by the assessment, comparing assessment results of known groups, comparing assessment results before and after a learning activity, and correlating assessment results with other measures. Procedures for collecting evidence using each of these methods were described.

Assessment–criterion relationship considerations for obtaining validity evidence focus on predicting future performance (the criterion) based on current assessment results. Obtaining this type of evidence involves a predictive validation study. If the assessment results are to be used to estimate students' performance on another assessment (the criterion measure) at the same time, the evidence is concurrent, and obtaining this type of evidence requires a concurrent validation study. Teachers rarely study the correlation of their own assessment results with criterion measures, but for tests with high-stakes outcomes, such as licensure and certification, this type of validity evidence is critical.

Ultimately, assessment validity requires an evaluation of interpretations and use of assessment results. The concept of validity thus has expanded to include *consideration of the consequences of assessment use* and how results are interpreted to students, teachers, and other stakeholders. Consideration of consequences must include both intended and unintended effects of assessment, particularly when assessment results are used to make high-stakes decisions.

A number of factors affect the validity of assessment results, including characteristics of the assessment itself, the administration and scoring procedures, and the test-takers. Each of these factors was discussed in some detail.

Reliability refers to the consistency of scores. Each assessment produces a limited measure of performance at a specific time. If this measurement is reasonably consistent over time, with different raters, or with different samples of the same domain, teachers can be more confident in the assessment results. Many extraneous factors may influence the measurement of performance, including instability of the behavior being measured, different samples of tasks in each assessment, varying assessment conditions between assessments, and inconsistent scoring procedures. These and other factors introduce error into every measurement. Methods of determining

assessment reliability estimate how much measurement error is present under varying assessment conditions. When assessment results are reasonably consistent, there is less measurement error and greater reliability.

Several points are important to an understanding of the concept of assessment reliability. Reliability pertains to assessment results, not to the assessment instrument itself. A reliability estimate always refers to a particular type of consistency, and it is possible for assessment results to be reliable in one or more of these respects but not in others. A reliability estimate always is calculated with statistical indices that express the relationship between two or more sets of scores. Reliability is an essential but insufficient condition for validity; low reliability always produces a low degree of validity, but a high reliability estimate does not guarantee a high degree of validity. Each of these points was discussed in this chapter.

Because reliability is viewed in terms of different types of consistency, these types are determined by different methods: over time (stability), among different forms of the assessment (equivalence), within the assessment itself (internal consistency), and among different raters (consistency of ratings or interrater reliability). Measures of stability indicate whether students would achieve essentially the same scores if they took the same assessment at another time—a test–retest procedure. Measures of equivalence involve the use of two or more forms of the same assessment, based on the same set of specifications (equivalent or alternate forms). Both forms of the assessment are administered to the same group of students in close succession, and the resulting scores are correlated. A high reliability coefficient indicates that teachers can make valid generalizations about student performance from one assessment to the other. Equivalent-form estimates of reliability are widely used in standardized testing, but are not practical for teacher-constructed assessments. Measures of internal consistency (split-half or half-length methods) can be used with a set of scores from only one administration of a single assessment. Estimates of internal consistency reveal the extent to which consistent results are obtained from two halves of the same assessment, revealing the extent to which the test items are internally consistent or homogeneous. Measures of consistency of ratings determine the extent to which ratings from two or more equally qualified persons agree on the score or rating. Interrater consistency is important to ensure that differences in stringency or leniency of ratings between raters do not place some students at a disadvantage. Use of scoring rubrics and training of raters to use those rubrics facilitates consensus among raters.

Various factors can influence the reliability of a set of test scores. These factors can be categorized into three main sources: the assessment instrument itself, the student, and the assessment administration conditions. Assessment-related factors include the length of the assessment, the homogeneity of assessment content, and the difficulty and discrimination ability of the individual items. Student-related factors

include the heterogeneity of the student group, test-taking ability, and motivation. Factors related to assessment administration include inadequate time to complete the test and cheating.

In addition, assessment tools should be practical and easy to use. Although reliability and validity are used to describe the ways in which scores are interpreted and used, practicality or usability is a quality of the instrument itself and its administration procedures. Assessment procedures should be efficient and economical. Teachers need to evaluate the following factors: ease of construction and use; time needed to administer and score the assessment and interpret the results; costs associated with assessment construction, administration, and scoring; and the ease with which assessment results can be interpreted simply and accurately by those who will use them.

■ References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Bosher, S. D. (2009). Removing language as a barrier to success on multiple-choice exams. In S. D. Bosher & M. D. Pharris (Eds.), *Transforming nursing education: The culturally inclusive environment* (pp. 259–284). New York, NY: Springer Publishing Company.
- Bosher, S. D., & Bowles, M. (2008). The effects of linguistic modification on ESL students' comprehension of nursing course test items. *Nursing Education Perspectives*, 29, 165–172.
- Brookhart, S. M., & Nitko, A. J. (2019). *Educational assessment of students* (8th ed.). New York, NY: Pearson Education.
- Gaberson, K. B., & Oermann, M. H. (2018). *Clinical teaching strategies in nursing education* (5th ed.). New York, NY: Springer Publishing Company.
- Goodwin, L. D. (1997). Changing conceptions of measurement validity. *Journal of Nursing Education*, 36, 102–107.
- Goodwin, L. D., & Goodwin, W. L. (1999). Measurement myths and misconceptions. *School Psychology Quarterly*, 14, 408–427.
- Miller, M. D., Linn, R. L., & Gronlund, N. E. (2013). *Measurement and assessment in teaching* (11th ed.). Upper Saddle River, NJ: Pearson Education.



TESTING AND OTHER ASSESSMENT METHODS

PLANNING FOR TESTING

It was Wednesday, and Paul Johnson was caught by surprise when he looked at his office calendar and realized that a test for the course he was teaching was only 1 week away, even though he was the person who had scheduled it! Thankful that he was not teaching this course for the first time, he searched his files for the test he had used last year. When he found it, his brief review showed that some of the content was outdated and that the test did not include items on the new content he had added this year. Because of a university policy that requires a minimum of 3 business days for the copy center to reproduce a test, Paul realized that he would have to finish the necessary revisions of the test and submit it for copying no later than Friday. He would be teaching in the clinical area on Thursday and teaching a class on Friday morning, and he was preparing to go out of town to attend a conference on Saturday.

He stayed up late on Wednesday night to revise the test, planning to proofread it on Thursday after he finished his clinical teaching responsibilities. But because of a family emergency, he was not able to proofread the test that night. Trusting that he had not made any serious clerical errors, he sent the test to the copy center before his class on Friday. When he returned to the office after his conference on Tuesday, he discovered that the copy center had been flooded during a severe storm over the weekend, and none of the photocopy jobs could be completed, including his test. Paul printed another copy of the test but could not take it anywhere else to be copied that day because of a scheduled committee meeting. To complicate matters, the department secretary had called in sick that day, and Paul could not change his childcare arrangements to allow him to stay late at the office to finish copying the test on the department machine. He came in very early on Wednesday morning to use the department photocopier, and finally finished the job just before the test was scheduled to begin.

With 5 minutes to spare, Paul rushed into the classroom and distributed the still-warm test booklets. As he was congratulating himself for meeting his deadline, the first student raised a hand with a question: "On item three, is there a typo?" Then another student said, "I don't think that the correct answer for item six is there." A third student complained, "Item nine is missing; the numbers jump from 8 to 10"

and a fourth student stated, “There are two ds for item 10.” Paul knew that it was going to be a long morning. But the worst was yet to come. As they were turning in their tests, students complained, “This test didn’t cover the material that I thought it would cover,” and “We spent a lot of class time analyzing case studies, but we were tested on memorization of facts.” Needless to say, Paul did not look forward to the posttest discussion the following week.

Too often, teachers give little thought to the preparation of their tests until the last minute and then rush to get the job done. A test that is produced in this manner often contains items that are poorly chosen, ambiguous, and either too easy or too difficult, as well as grammatical, spelling, and other clerical errors. The solution lies in adequate planning for test construction before the item-writing phase begins, followed by careful critique of the completed test by other teachers. Exhibit 3.1 lists the steps of the test-construction process. This chapter describes the steps involved in planning for test construction; subsequent chapters will focus on the techniques of writing test items of various formats, assembling and administering the test, and analyzing the test results.

EXHIBIT 3.1

CHECKLIST FOR TEST CONSTRUCTION

- ☐ Define the purpose of the test.
- ☐ Describe the population to be tested.
- ☐ Determine the optimum length of the test.
- ☐ Specify the desired difficulty and discrimination levels of the test items.
- ☐ Determine the scoring procedure or procedures to be used.
- ☐ Select item formats to be used.
- ☐ Construct a test blueprint or table of specifications.
- ☐ Write the test items.
- ☐ Have the test items critiqued.
- ☐ Determine the arrangement of items on the test.
- ☐ Write specific directions for each item format.
- ☐ Write general directions for the test and prepare a cover sheet.
- ☐ Print or type the test.
- ☐ Proofread the test.
- ☐ Reproduce the test.
- ☐ Prepare a scoring key.
- ☐ Prepare students for taking the test.

■ Purpose and Population

All decisions involved in planning a test are based on a teacher's knowledge of the purpose of the test and the relevant characteristics of the population of learners to be tested. The *purpose* for the test involves why it is to be given, what it is supposed to measure, and how the test scores will be used. For example, if a test is to be used to measure the extent to which students have met learning objectives to determine course grades, its primary purpose is summative. If the teacher expects the course grades to reflect real differences in the amount of knowledge among the students, the test must be sufficiently difficult to produce an acceptable range of scores. On the other hand, if a test is to be used primarily to provide feedback to staff nurses about their knowledge following a continuing education program, the purpose of the test is formative. If the results will not be used to make important personnel decisions, a large range of scores is not necessary, and the test items can be of moderate or low difficulty. In this chapter, the outcomes of learning are referred to as *objectives*, but as discussed in Chapter 1, Assessment and the Educational Process, many nurse educators refer to these as *outcomes*. A teacher's knowledge of the population that will be tested will be useful in selecting the item formats to be used, determining the length of the test and the testing time required, and selecting the appropriate scoring procedures. The term *population* is not used here in its research sense, but rather to indicate the general group of learners who will be tested. The students' reading levels, English-language literacy, visual acuity, health, and previous testing experience are examples of factors that might influence these decisions. For example, if the population to be tested is a group of five patients who have completed preoperative instruction for coronary bypass graft surgery, the teacher would probably not administer a test of 100 multiple-choice and matching items with a machine-scored answer sheet. However, this type of test might be most appropriate as a final course examination for a class of 75 senior nursing students.

■ Test Length

The length of the test is an important factor that is related to its purpose, the abilities of the students, the item formats to be used, the amount of testing time available, and the desired reliability of the test scores. As discussed in Chapter 2, Qualities of Effective Assessment Procedures: Reliability, Validity, and Usability, the reliability of test scores generally improves as the length of the assessment increases, so the teacher should attempt to include as many items as possible to adequately sample the content. However, if the purpose of the test is to measure knowledge of a small content domain with a limited number of objectives, fewer items will be needed to achieve an adequate sampling of the content.

It should be noted that *assessment length* refers to the number of test items or tasks, not to the amount of time it would take the student to complete the test. Items that require the student to analyze a complex data set, draw conclusions, and supply or choose a response take more test administration time; therefore, fewer items of those types can be included on a test to be completed in a fixed time period. When the number of complex assessment tasks to be included on a test is limited by test administration time, it is better to test more frequently than to create longer tests that test less important learning goals (Brookhart & Nitko, 2019; Miller, Linn, & Gronlund, 2013).

Because test length probably is limited by the scheduled length of a testing period, it is wise to construct the test so that the majority of the students working at their normal pace will be able to attempt to answer all items. This type of test is called a *power* test. A *speeded* test is one that does not provide sufficient time for all students to respond to all items. Although most standardized tests are speeded, this type of test generally is not appropriate for teacher-made tests in which accuracy rather than speed of response is important (Brookhart & Nitko, 2019; Miller et al., 2013).

■ Difficulty and Discrimination Level

The desired difficulty of a test and its ability to differentiate among various levels of performance are related considerations. Both factors are affected by the purpose of the test and the way in which the scores will be interpreted and used. The difficulty of individual test items affects the average test score; the mean score of a group of students is equal to the sum of the difficulty levels of the test items. The difficulty level of each test item depends on the complexity of the task, the ability of the students who answer it, and the quality of the teaching. It also may be related to the perceived complexity of the item; if students perceive the task as too difficult, they may skip it, resulting in a lower percentage of students who answer the item correctly (Brookhart & Nitko, 2019). See Chapter 12, *Test and Item Analysis: Interpreting Test Results*, for a more detailed discussion of item difficulty and discrimination.

If test results are to be used to determine the relative achievement of students (i.e., norm-referenced interpretation), the majority of items on the test should be moderately difficult. The recommended difficulty level for selection-type test items depends on the number of choices allowed. The percentage of students who answer each item correctly should be about midway between 100% and the chance of guessing correctly (e.g., 50% for true-false items, 25% correct for four-alternative multiple-choice items). For example, a moderately difficult true-false item should be answered correctly by 75% to 85% of students (Waltz, Strickland, & Lenz, 2010). When the majority of items on a test are too easy or too difficult, they will not discriminate well between students with varying levels of knowledge or ability.

However, if the teacher wants to make criterion-referenced judgments, more commonly used in nursing education and practice settings, the overall concern is

whether a student's performance meets a set standard rather than on the actual score itself. If the purpose of the assessment is to screen out the least capable students (e.g., those failing a course), it should be relatively easy for most test-takers. However, comparing performance to a set standard does not limit assessment to testing of lower level knowledge and ability; considerations of assessment validity should guide the teacher to construct tests that adequately sample the knowledge or performance domain.

When criterion-referenced test results are reported as percentage scores, their variability (range of scores) may be similar to norm-referenced test results, but the interpretation of the range of scores would be more narrow. For example, on a final exam in a nursing course, the potential score range may be 0% to 100%, but the passing score is set at 80%. Even if there is wide variability of scores on the exam, the primary concern is whether the test correctly classifies each student as performing above or below the standard (e.g., 80%). In this case, the teacher should examine the difficulty level of test items and compare them between groups (students who met the standard and students who did not). If item difficulty levels indicate a relatively easy or relatively difficult exam, criterion-referenced decisions will still be appropriate if the measure consistently classifies students according to the performance standard (Miller et al., 2013; Waltz et al., 2010).

It is important to keep in mind that the difficulty level of test items can only be estimated in advance, depending on the teacher's experience in testing this content and knowledge of the abilities of the students to be tested. When the test has been administered and scored, the actual difficulty index for each item can be compared with the expected difficulty, and items can be revised if the actual difficulty level is much lower or much higher than anticipated (Waltz et al., 2010). Procedures for determining how the test items actually perform are discussed in Chapter 12, *Test and Item Analysis: Interpreting Test Results*.

■ Item Formats

Some students may be particularly adept at answering essay items; others may prefer multiple-choice items. However, tests should be designed to provide information about students' knowledge or abilities, not about their skill in taking certain types of tests. A test with a variety of item formats provides students with multiple ways to demonstrate their competence (Brookhart & Nitko, 2019). All item formats have their advantages and limitations, which are discussed in later chapters.

Selection Criteria for Item Formats

Teachers should select item formats for their tests based on a variety of factors, such as the learning outcomes to be evaluated, the specific skill to be measured, and the

ability level of the students. Some objectives are better measured with certain item formats. For example, if the instructional objective specifies that the student will be able to “discuss the comparative advantages and disadvantages of breast and bottle-feeding,” a multiple-choice item would be inappropriate because it would not allow the teacher to evaluate the student’s ability to organize and express ideas on this topic. An essay item would be a better choice for this purpose. Essay items provide opportunities for students to formulate their own responses, drawing on prior learning, and to express their ideas in writing; these often are desired outcomes of nursing education programs.

The teacher’s time constraints for constructing the test may affect the choice of item format. In general, essay items take less time to write than multiple-choice items, but they are more difficult and time-consuming to score. A teacher who has little time to prepare a test and therefore chooses an essay format, assuming that this choice is also appropriate for the objectives to be tested, must plan for considerable time after the test is given to score it.

In nursing education programs, faculty members often develop multiple-choice items as the predominant, if not exclusive, item format because for a number of years, licensure and certification examinations contained only multiple-choice items. Although this type of test item provides essential practice for students in preparation for taking such high-stakes examinations, it contradicts the principle of selecting the most appropriate type of test item for the outcome and content to be evaluated. In addition, it limits variety in testing and creativity in evaluating student learning. Although practice with multiple-choice items questions is critical, other types of test items and evaluation strategies also are appropriate for measuring student learning in nursing. In fact, although many of the NCLEX® (National Council Licensure Examination) examination items are four-option multiple-choice items, the item pools now contain other formats such as completion and multiple response (National Council of State Boards of Nursing, 2019). Nurse educators should not limit their selection of item formats based on the myth that learners must be tested exclusively with the item format most frequently used on a licensure or certification test.

On the other hand, each change of item format on a test requires a change of task for students. Therefore, the number of different item formats to include on a test also depends on the length of the test and the level of the learner. It is generally recommended that teachers use no more than three item formats on a test. Shorter assessments, such as a 10-item quiz, may be limited to a single item format.

Objectively and Subjectively Scored Items

Another powerful and persistent myth is that some item formats evaluate students more objectively than do other formats. Although it is common to describe true-false, matching, and multiple-choice items as “objective,” *objectivity* refers to the

way items are scored, not to the type of item or their content (Miller et al., 2013). Objectivity means that once the scoring key is prepared, it is possible for multiple teachers on the same occasion or the same teacher on multiple occasions to arrive at the same score. Subjectively scored items, like essay items (and short-answer items, to a lesser extent), require the judgment of the scorer to determine the degree of correctness and therefore are subject to more variability in scoring.

Selected-Response and Constructed-Response Items

Another way of classifying test items is to identify them by the type of response required of the test-taker (Miller et al., 2013; Waltz et al., 2010). *Selected-response* (or “choice”) items require the test-taker to select the correct or best answer from among options provided by the teacher. In this category, the item formats are true–false, matching exercises, and multiple-choice. *Constructed-response* (or “supply”) formats require the learner to supply an answer, and may be classified further as limited response (or short response) and extended response. These are the short-answer and essay formats. Exhibit 3.2 depicts this schema for classifying test-item formats and the variations of each type.

■ Scoring Procedures

Decisions about what scoring procedure or procedures to use are somewhat dependent on the choice of item formats. Student responses to short-answer, numerical calculation, and essay items, for instance, usually must be hand-scored, whether they are recorded directly on the test itself, on a separate answer sheet, or in a booklet. Answers to objective test items such as multiple choice, true–false, and matching also may be recorded on the test itself or on a separate answer sheet. Scannable answer sheets greatly increase the speed of objective scoring procedures and have the additional advantage of allowing computer-generated item analysis reports to be produced. The teacher should decide if the time and resources

EXHIBIT 3.2

CLASSIFICATION OF TEST ITEMS BY TYPE OF RESPONSE

SELECTED-RESPONSE ITEM FORMATS ("CHOICE" ITEMS)	CONSTRUCTED-RESPONSE ITEM FORMATS ("SUPPLY" ITEMS)
True–false	Short answer
Matching exercises	Completion or fill in the blank
Multiple choice	Restricted-response essay
Multiple response	Extended-response essay

available for scoring a test suggest that hand scoring or electronic scoring would be preferable. In any case, this decision alone should not influence the choice of test-item format.

■ Test Blueprint

Most people would not think of building a house without blueprints. In fact, the word *house* denotes diverse attributes to different individuals. For this reason, a potential homeowner would not purchase a lot, call a builder, and say only, “Build a house for me on my lot.” The builder might think that a proper house consists of a two-story brick colonial with four bedrooms, three baths, and a formal dining room, whereas the homeowner had a three-bedroom ranch with two baths, an eat-in kitchen, and a great room with a fireplace in mind. Similarly, the word *test* might mean different things to different people; students and their teacher might have widely varied expectations about what the test will contain. The best way to avoid misunderstandings regarding the nature of a test and to ensure that the teacher will be able to make valid judgments about the test scores is to develop a test blueprint, also known as a *test plan* or a *table of specifications*, before “building” the test itself.

The elements of a test blueprint include (a) a list of the major topics or instructional objectives (or both) that the test will cover; (b) the types of thinking skills or level of complexity of the task to be assessed; and (c) the emphasis each topic will have, indicated by number or percentage of items or points (Brookhart & Nitko, 2019). Exhibit 3.3 is an example of a test blueprint for a unit test on nursing care during normal pregnancy that illustrates each of these elements.

The row headings along the left margin of the example are the content areas that will be tested. In this case, the content is indicated by a general outline of topics. Teachers may find that a more detailed outline of content or a list of the relevant objectives is more useful for a given purpose and population. Some teachers combine a content outline and the related objectives; in this case, an additional column of objectives would be inserted before or after the content list.

The column headings across the top of the example are taken from the taxonomy of cognitive domain (Anderson & Krathwohl, 2001). Because the test blueprint is a tool to be used by the teacher, it can be modified in any way that makes sense to the user. Accordingly, the teacher who prepared this blueprint chose to use only selected levels of the taxonomy. Other teachers might include all levels or different levels of the taxonomy, or use a different taxonomy.

The body of the test blueprint is a grid formed by the intersections of content topics and cognitive levels. Each of the cells of the grid has the potential to represent one or more test items that might be developed. The numbers in the cells of the

EXHIBIT 3.3**EXAMPLE OF A TEST BLUEPRINT FOR A UNIT TEST ON NORMAL PREGNANCY (75 POINTS)**

CONTENT	LEVEL OF COGNITIVE SKILL ^a				
	R	U	Ap	An	TOTAL # ^b
I. Conception and fetal development		2	3	3	8
II. Maternal physiological changes in pregnancy	2	3	1	2	8
III. Maternal psychological changes in pregnancy		2	2	3	7
IV. Social, cultural, and economic factors affecting pregnancy outcome		3	2	3	8
V. Signs and symptoms of pregnancy	2	2	2		6
VI. Antepartal nursing care		8	10	12	30
VII. Preparation for childbirth		4	1	3	8
TOTAL # ^b	4	24	21	26	75

Note: ^aAccording to Anderson and Krathwohl (2001) taxonomy of the cognitive domain. Selected levels are included in this test blueprint and are represented by the following key:
R = Remembering
U = Understanding
Ap = Applying
An = Analyzing
^bNumber of points. Test blueprints also may include the number or the percentage of items.

sample test blueprint represent the number of points on the test that will relate to it; some teachers prefer to indicate numbers of items or the percentage of points or items represented by each cell. The percentage is a better indicator of the amount of emphasis to be given to each content area (Miller et al., 2013), but the number of items or points may be more helpful to the teacher in writing actual test items. Students usually expect that the variation in number of points allotted to each content area or objective relates to the amount of time devoted to it in class or to the emphasis that the teacher placed on it. A cell value can represent the number of items on a test in which each item is worth 1 point (such as multiple choice), or it can indicate one multipoint item (such as an essay item), or any combination of those

items (Brookhart & Nitko, 2019). It is not necessary to write test items for each cell; the teacher's judgment concerning the appropriate emphasis and balance of content governs the decision about which cells should be filled and how many items should be written for each.

Rigorous classification of items into these cells also is unnecessary and, in fact, impossible; the way in which the content is actually taught may affect whether the related test items will be written at the applying or understanding level, for example. For this reason, the actual test may deviate slightly from the specifications for certain cells, but the overall balance of emphasis between the test and the actual instruction should be very similar (Brookhart & Nitko, 2019; Miller et al., 2013).

Once developed, the test blueprint serves several important functions. First, it is a useful tool for guiding the work of the item writer so that sufficient items are developed at the appropriate level to test important content areas and objectives. Without a test blueprint, teachers often use ease of construction as a major consideration in writing test items, resulting in tests with a limited and biased sample of learning tasks that may omit outcomes of greater importance that are more difficult to measure (Miller et al., 2013). Using test blueprints also helps teachers to be accountable for the educational outcomes they produce. The test blueprint can be used as evidence for judging the validity of the resulting test scores. The completed test and blueprint may be reviewed by content experts who can judge whether the test items adequately represent the specified content domain, as described in the procedures for collecting content-related evidence in Chapter 2, *Qualities of Effective Assessment Procedures: Reliability, Validity, and Usability*.

Another important use of the test blueprint is to inform students about the nature of the test and how they should prepare for it. Although the content covered in class and assigned readings should give students a general idea of the content areas to be tested, students often lack a clear sense of the cognitive levels at which they will be tested on this material. Although it might be argued that the objectives might give students a clue as to the level at which they will be tested, teachers often forget that students are not as sophisticated in interpreting objectives as teachers are. Also, some teachers are good at writing objectives that specify a reasonable expectation of performance, but their test items may in fact test higher or lower performance levels. Students need to know the level at which they will be tested because that knowledge will affect how they prepare for the test, not necessarily how much they prepare. They should prepare differently for items that test their ability to apply information than for items that test their ability to synthesize information.

Some teachers worry that if the test blueprint is shared with students, they will not study the content areas that would contribute less to their overall test scores, preferring to concentrate their time and energy on the more important areas of emphasis. If this indeed is the outcome, is it necessarily harmful? Lacking any guidance from the

teacher, students may unwisely spend equal amounts of time reviewing all content areas. In fact, professional experience reveals that some knowledge is more important for use in practice than other knowledge. Even if they are good critical thinkers, students may be unable to discriminate more important content from that which is less important because they lack the practice experience to make this distinction. Withholding information about the content emphasis of the test from students might be perceived as an attempt to threaten or punish them for perceived shortcomings such as failure to attend class, failure to read what was assigned, or failure to discern the teacher's priorities. Such a use of testing would be considered unethical.

The best time to share the test blueprint with students is at the beginning of the course or unit of study. If students are unfamiliar with the use of a test blueprint, the teacher may need to explain the concept as well as discuss how it might be useful to the students in planning their preparation for the test. Of course, if the teacher subsequently makes modifications in the blueprint after writing the test items, those changes also should be shared with the students (Brookhart & Nitko, 2019).

■ Writing the Test Items

After developing the test blueprint, the teacher should begin to write the test items that correspond to each cell. Regardless of the selected item formats, the teacher should consider some general factors that contribute to the quality of the test items.

General Rules for Writing Test Items

1. *Every item should measure something important.* If a test blueprint is designed and used as described in the previous section, each test item will measure an important objective or content area. Without using a blueprint, teachers often write test items that test trivial or obscure knowledge. Sometimes the teacher's intent is to determine whether the students have read assigned materials; however, if the content is not important information, it wastes the teacher's time to write the item and wastes the students' time to read it and respond to it. Similarly, it is not necessary to write "filler" items to meet a targeted number; a test with 98 well-written items that measure important objectives will work as well as or better than one with 98 good items and two meaningless ones. Although the reliability of test results is related to the length of the assessment, this rule presumes that items added to a test to increase the number of tasks would be of the same quality as those that are already a part of the test. Adding items that are so easy that every student will answer the questions correctly, or so difficult that every student will answer them incorrectly, will not improve the reliability estimate (Miller et al., 2013). In fact, students who know the content well might regard a test item that measures trivial knowledge with

annoyance or even suspicion, believing that it is meant to trick them into answering incorrectly. There is no reason other than ease of mentally calculating a percentage score for setting an absolute target number of points on a test at 100.

2. *Every item should have a correct answer.* The correct answer should be one that would be agreed on by experts (Miller et al., 2013). This may seem obvious, but the rule is violated frequently because of the teacher's failure to make a distinction between fact and belief. In some cases, the correct or best answer to a test item might be a matter of opinion, and unless a particular authority is cited in the item, students might justifiably argue a different response than the one the teacher expected. For example, one answer to the question, "When does life begin?" might be "When the kids leave home and the dog dies." If the intent of the question was to measure understanding of when a fetus becomes viable, this is not the correct answer, although if the latter was the teacher's intent, the question was poorly worded. There are a variety of opinions and beliefs about the concept of viability; a better way to word this question is, "According to the standards of the American College of Obstetricians and Gynecologists, at what gestational age does a fetus become viable?" If a test item asks the student to state an opinion about an issue and to support that position with evidence, that is a different matter. That type of item should not be scored as correct or incorrect, but with variable credit based on the completeness of the response, rationale given for the position taken, or the soundness of the student's reasoning (Brookhart & Nitko, 2019).
3. *Use simple, clear, concise, precise, grammatically correct language.* Students who read the test item need to know exactly what task is required of them. Wording a test item clearly is often difficult because of the inherent abstractness and imprecision of language, and it is a challenge to use simple words and sentence structure when writing about highly technical and complex material. The teacher should include enough detail in the test item to communicate the intent of the item but without extraneous words or complex syntax that only serve to increase the reading time. In addition, grammatical errors may provide unintentional clues to the correct response for the testwise but unprepared student and, at best, annoy the well-prepared student.

This rule is particularly important when testing students who are non-native speakers (NNSs). Boshier and Bowles (2008) found that in a majority of cases, linguistic modification of test items improved NNSs' comprehension of nursing exam items. The process of linguistic modification or simplification maintains key content area vocabulary but reduces the semantic and syntactic complexity of written English. Linguistic structures such as passive voice constructions, long question phrases, conditional and subordinate clauses, negation, and grammatical errors are particularly difficult for NNSs to understand, and they require more time to read and process (Boshier, 2009; Boshier &

Bowles, 2008). Although arguments might be made that no accommodation is made for NNSs on the NCLEX, consideration of measurement validity must take into account that any test that employs language is at least partially a measure of language skills (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; Miller et al., 2013).

The following item stems, adapted from an example given by Boshier and Bowles (2008), illustrate the effect of linguistic simplification:

Original stem: A patient with chronic pain treated over a period of months with an oral form of morphine tells you that she is concerned because she has had to gradually increase the amount of medication she takes to achieve pain control. Your response should include:

Linguistically simplified stem: A patient has chronic pain. She is treated over a period of months with an oral form of morphine. She tells the nurse that she is concerned because she has gradually needed more medication to achieve the same level of pain control. How should the nurse respond? (Boshier & Bowles, 2008, p. 168)

Note that the same content is emphasized, but that the revised example contains four short simple sentences and ends with a question to be answered rather than a completion format. Given growing concerns that even native English speakers are entering postsecondary programs with poorer reading skills, such linguistic modification should benefit all students.

4. *Avoid using jargon, slang, or unnecessary abbreviations.* Healthcare professionals frequently use jargon, abbreviations, and acronyms in their practice environment; in some ways, it allows them to communicate more quickly, if not more effectively, with others who understand the same language. Informal language in a test item, however, may fail to communicate the intent of the item accurately. Because most students are somewhat anxious when taking tests, they may fail to interpret an abbreviation correctly for the context in which it is used. For example, does MI mean *myocardial infarction*, *mitral insufficiency*, or *Michigan*? Of course, if the intent of the test item is to measure students' ability to define commonly used abbreviations, it would be appropriate to use the abbreviation in the item and ask for the definition, or give the definition and ask the student to supply the abbreviation. Slang almost always conveys the impression that the item writer does not take the job seriously. As noted previously, slang, jargon, abbreviations, and acronyms contribute to linguistic complexity, especially for NNSs. In addition, growing alarm about healthcare errors attributed to poor communication, including the overuse of abbreviations, suggests that nurse educators should set positive examples for their students by using only abbreviations generally approved for use in clinical settings.

5. *Try to use positive wording.* It is difficult to explain this rule without using negative wording, but in general, avoid including words like *no*, *not*, and *except* in the test item. As noted previously, negation contributes to linguistic complexity that interferes with the test performance of NNSs. The use of negative wording is especially confusing in true–false items. If using a negative form is unavoidable, underline the negative word or phrase, or use bold text and all uppercase letters to draw students’ attention to it. It is best to avoid asking students to identify the incorrect response, as in the following example:

Which of the following is **NOT** an indication that a skin lesion is a Stage IV pressure ulcer?

- a. Blistering¹
- b. Sinus tracts
- c. Tissue necrosis
- d. Undermining

The structure of this item reinforces the wrong answer and may lead to confusion when a student attempts to recall the correct information at a later time. A better way to word the item is:

Which of the following is an indication that a skin lesion is a Stage II pressure ulcer?

- a. Blistering¹
- b. Sinus tracts
- c. Tissue necrosis
- d. Undermining

6. *No item should contain irrelevant clues to the correct answer.* This is a common error among inexperienced test-item writers. Students who are good test-takers can usually identify such an item and use its flaws to improve their chances of guessing the correct answer when they do not know it. Irrelevant clues include a multiple-choice stem that is grammatically inconsistent with one or more of the options, a word in the stem that is repeated in the correct option, using qualifiers such as “always” or “never” in incorrect responses, placing the correct response in a consistent position among a set of options, or consistently making true statements longer than false statements (Brookhart & Nitko, 2019; Miller et al., 2013). Such items contribute little to the validity of test results because they may not measure what students actually know, but how well they are able to guess the correct answers.
7. *No item should depend on another item for meaning or for the correct answer.* In other words, if a student answers one item incorrectly, he or she will likely

¹Correct answer.

answer the related item incorrectly. An example of such a relationship between two completion items follows:

1. Which insulin should be used for emergency treatment of ketoacidosis?

2. What is the onset of action for the insulin in Item 1?

In this example, Item 2 is dependent on Item 1 for its meaning. Students who supply the wrong answer to Item 1 are unlikely to supply a correct answer to Item 2. Items should be worded in such a way as to make them independent of each other. However, a series of test items can be developed to relate to a context such as a case study, database, diagram, graph, or other interpretive material. Items that are linked to this material are called *interpretive* or *context-dependent* items, and they do not violate this general rule for writing test items because they are linked to a common stimulus, not to each other.

8. *Eliminate extraneous information unless the purpose of the item is to determine whether students can distinguish between relevant and irrelevant data.* Avoid the use of patient names in clinical scenarios; this information adds unnecessarily to reading time, it may distract from the purpose of the item, and it may introduce cultural bias (see Chapter 16, Social, Ethical, and Legal Issues). However, some items are designed to measure whether a student can evaluate the relevance of clinical data and use only pertinent information in arriving at the answer. In this case, extraneous data (but not patient names) may be included.
9. *Arrange for a critique of the items.* The best source of this critique is a colleague who teaches the same content area or at least someone who is skilled in the technical aspects of item writing. If no one is available to critique the test items, the teacher who developed them should set them aside for a few days. This will allow the teacher to review the items with a fresh perspective to identify lack of clarity or faulty technical construction.
10. *Prepare more items than the test blueprint specifies.* This will allow for replacement items for those discarded in the review process. The fortunate teacher who does not need to use many replacement items can use the remainder to begin an item bank for future tests.

■ Preparing Students to Take a Test

A teacher-made test usually measures students' maximum performance rather than their typical performance. For this reason, teachers should create conditions under

which students will be able to demonstrate their best possible performance. These conditions include adequate preparation of students to take the test (Brookhart & Nitko, 2019; Miller et al., 2013). Although this is the last point on the test-construction checklist (Exhibit 3.1), the teacher should begin preparing students to take the test at the time the test is scheduled. Adequate preparation includes information, skills, and attitudes that will facilitate students' maximum performance on the test.

Information Needs

Students need information about the test to plan for effective preparation. They need sufficient time to prepare for a test, and the date and time of a test should be announced well in advance. Although many teachers believe that unannounced or “pop” tests motivate students to study more, there is no evidence to support this position. In fact, surprise (unscheduled) tests can be considered punitive or threatening and, as such, represent an unethical use of testing (Brookhart & Nitko, 2019). Adult learners with multiple responsibilities may need to make adjustments to their work and family responsibilities to have adequate study time, and generous notice of a planned test date will allow them to set their priorities.

In addition, students need to know about the conditions under which they are to be tested, such as how much time will be allotted, whether they will have access to resources such as textbooks, how many items will be included, the types of item formats that will be used, and whether they need special tools or supplies to take the test, such as calculators, pencils, or black-ink pens (Miller et al., 2013). They also should know what items and resources they will not be able to use during the test. For example, the teacher may direct students not to bring cell phones, personal digital assistants, chiming watches, watches with calculators, backpacks, briefcases, or any books or papers to the testing site. Some teachers do not allow students to wear caps or hats with brims to discourage cheating. In fact, such requirements may be good practice for prelicensure students who must observe similar restrictions for the NCLEX.

Of course, students also should know what content will be covered on the test, how many items will be devoted to each content area, the cognitive level at which they will be expected to perform, and the types of items to expect. As previously discussed, giving students a copy of the test blueprint and discussing it with them is an effective way for teachers to convey this information. Students should also have sufficient opportunity to practice the type of performance that will be tested. For example, if students will be expected to solve medication dose calculation problems without the use of a calculator, they should practice this type of calculation in class exercises or out-of-class assignments. Students also need to know whether spelling, grammar, punctuation, or organization will be considered in scoring open-ended items so that they can prepare accordingly. Finally, teachers should tell students how

their test results will be used, including the weight assigned to the test score in grading (Brookhart & Nitko, 2019; Miller et al., 2013).

Another way that teachers can assist students in studying for a test is to have students prepare and use a “cheat sheet.” Although this term can be expected to have negative connotations for most teachers, cheat sheets commonly are used in nursing practice in the form of memory aids or triggers such as procedure checklists, pocket guides, and reminder sheets. When legitimized for use in studying and test-taking, cheat sheets capitalize on the belief that although dishonest behavior must be discouraged, the skills associated with cheating can be powerful learning tools.

When students intend to cheat on a test, they usually try to guess potential test items and prepare cheat sheets with the correct answers to those anticipated items. Using this skill for a more honest purpose, the teacher can encourage all students to anticipate potential test items. In a test-preparation context, the teacher requires the students to develop a written cheat sheet that summarizes, prioritizes, condenses, and organizes content that they think is important and wish to remember during the test. The teacher may set parameters such as the length of the cheat sheet—for example, one side of one sheet of $8\frac{1}{2} \times 11$ -inch paper. The students bring their cheat sheets on the day of the test and may use them during the test; they submit their cheat sheets along with their test papers. Students who do not submit cheat sheets may be penalized by deducting points from their test scores or may not be permitted to take the test at all.

Some students may not even consult their cheat sheets during the test, but they still derive benefit from the preparation that goes into developing them. The teacher also may review the cheat sheets with students whose test scores are low to identify weaknesses in thinking that may have contributed to their errors. When used for this purpose, the cheat sheet becomes a powerful diagnostic and feedback tool.

Test-Taking Skills

Because of an increasingly diverse population of learners in every educational setting, including growing numbers of students for whom English is a second language and whose testing experiences may be different from the teacher’s expectations, teachers should determine whether their students have adequate test-taking skills for the type of test to be given. If the students lack adequate test-taking skills, their test scores may be lower than their actual abilities. Skill in taking tests sometimes is called *testwiseness*. To be more precise, testwiseness is the ability to use test-taking skills, clues from poorly written test items, and test-taking experience to achieve a test score that is higher than the student’s true knowledge would predict. Common errors made by item writers do allow some students to substitute testwiseness for knowledge. But, in general, all students should develop adequate test-taking skills so that they are not at a disadvantage when their scores are compared with those of

more testwise individuals. Adequate test-taking skills include the following abilities (Brookhart & Nitko, 2019):

1. Reading and listening to directions and following them accurately
2. Reading test items carefully
3. Recording answers to test items accurately and neatly
4. Avoiding physical and mental fatigue by paced study and adequate rest before the test rather than late-night cram sessions supplemented by stimulants
5. Using test time wisely and working at a pace that allows for careful reflection but also permits responding to all items that the student is likely to answer correctly
6. Bypassing difficult items and returning to them later
7. Making informed guesses rather than omitting answers
8. Outlining and organizing responses to essay items before beginning to write
9. Checking answers to test items for clerical errors and changing answers if a better response is indicated

Many teachers advise students not to change their answers to test items, believing that the first response usually is the correct answer and that changing responses will not increase a student's score. Research findings, however, do not support this position. Studies of answer changing and its effect on test performance have revealed that most students do change their answers to about 4% of test items and that approximately two thirds of answer changes become correct responses. As item difficulty increases, however, this payoff diminishes; consequently, more knowledgeable students benefit more than less knowledgeable students from changing answers (Brookhart & Nitko, 2019).

Students should be encouraged to change their first response to any item when they have a good reason for making the change. For example, a student who has a clearer understanding of an item after rereading it, who later recalls additional information needed to answer the item, or who receives a clue to the correct answer from another item should not hesitate to change the first answer. Improvement in test scores should not be expected, however, when students change answers without a clear rationale for making the change.

Test Anxiety

Finally, teachers should prepare students to approach a test with helpful attitudes. Although anxiety is a common response to situations in which performance is evaluated, high levels of anxiety are likely to interfere with maximum performance (Miller et al., 2013).

Whether some students can be characterized as test-anxious is a matter of frequent debate. Test anxiety can be viewed in several ways. Students who are motivated to do well often experience increased emotional tension in response to a test. Their perceptions of the testing situation affect their thoughts during test preparation and test-taking. Students who perceive a test as a challenge usually have thoughts that are task-directed. They can focus on completing the task and easily manage any tension that is associated with it. Some students perceive tests as threats because they have poor test-taking skills, inadequate knowledge, or both. These students often have task-irrelevant thoughts about testing. They focus on what could happen if they fail a test, and their feelings of helplessness cause them to desire to escape the situation (Brookhart & Nitko, 2019).

Test anxiety can be characterized as a trait with three components: physical, emotional, and cognitive. Test-anxiety research suggests an interaction among these components: negative thoughts and perceptions about testing can create negative feelings, which interfere with performance (Poorman, Mastorovich, Molcan, & Liberto, 2011). The physical component, or autonomic reactivity, involves unpleasant feelings and reactions such as perspiration, increased heart rate, headaches, and gastrointestinal symptoms, although not all test-anxious individuals have physical reactions. The emotional component involves mood and feelings (e.g., nervousness, uneasiness, fear, dread, panic) associated with testing situations.

The cognitive component refers to thoughts or concerns related to performance and its consequences, occurring before or during a test. Essentially, the cognitive component involves worry about possible negative outcomes: “catastrophic fantasies” about what might happen if the student fails, and “competitive worry” that other students are doing better (Poorman et al., 2011). Cognitive indications of test anxiety include impaired ability to concentrate and easy distractibility during the test, difficulty recalling information (“going blank”), misreading or misunderstanding directions or test items, and feeling pressured to be perfect. In addition, individuals with true test anxiety often have a history of poor performance on tests and other evaluative situations, particularly high-stakes tests. For example, these individuals may repeatedly fail a driver’s license examination or achieve good scores on quizzes or unit tests but fail final examinations (Poorman et al., 2011).

The combination of negative feelings and thoughts often results in behaviors that interfere with students’ ability to prepare adequately for a test. One of the most dangerous behaviors is avoidance—procrastinating rather than beginning preparation early, and engaging in activities that seem to be related to preparing for the test but really are just distractions. For example, students often report that they studied for many hours and still failed a test, but a record of their activities would reveal that much of that time was spent highlighting material in the textbook or “preparing to study”—organizing their notes, doing household chores with the intention of

minimizing interruptions, and so on. Negative thinking creates anxiety, which students try to avoid by avoiding the studying that they believe is causing the discomfort (Poorman et al., 2011).

Students whose test anxiety interferes with their performance often benefit from treatment that addresses the feeling or emotional component of anxiety and the negative thinking or worry aspect as well as training to improve their general test-taking skills. For example, the test-anxious student may learn techniques for stopping negative thoughts during study periods and testing situations, and behavioral techniques such as progressive relaxation and visual imagery (Poorman et al., 2011). A more comprehensive discussion of the diagnosis and treatment of test anxiety is beyond the scope of this textbook. However, teachers may be able to identify students whose performance suggests that test anxiety may be a factor, and to work with them so that they perform at their best, or refer those students for treatment.

Students need to view tests and other assessment procedures as opportunities to demonstrate what they know and what they can do. To foster this attitude, the teacher should express confidence in the students' abilities to prepare for and perform well on an upcoming test. It may be helpful for the teacher to ask the students what would help them to feel more relaxed and less anxious before and during a test. Conducting a review session, giving practice items similar to those that will be used on the test, and not talking or interrupting students during a test are examples of strategies that are likely to reduce students' anxiety to manageable levels (Brookhart & Nitko, 2019; Miller et al., 2013).

■ Summary

Teachers who leave little time for adequate preparation often produce tests that contain poorly chosen and poorly written test items. Sufficient planning for test construction before the item-writing phase begins, followed by a careful critique of the completed test by other teachers, is likely to produce a test that will yield more valid results.

All decisions involved in planning a test should be based on a teacher's knowledge of the purpose of the test and relevant characteristics of the population of learners to be tested. The purpose for the test involves why it is to be given, what it is supposed to measure, and how the test scores will be used. A teacher's knowledge of the population that will be tested will be useful in selecting the item formats to be used, determining the length of the test and the testing time required, and selecting the appropriate scoring procedures. The students' English-language literacy, visual acuity, and previous testing experience are examples of factors that might influence these decisions.

The length of the test is an important factor that is related to its purpose, the abilities of the students, the item formats that will be used, the amount of testing time available, and the desired reliability of the test scores. The desired difficulty of the test and its ability to differentiate among various levels of performance are affected by the purpose of the test and the way in which the scores will be interpreted and used.

A test with a variety of item formats usually provides students with more opportunity to demonstrate their competence than a test with only one item format. Test items may be classified as selected-response or constructed-response types, depending on the task required of the learner. All item formats have advantages and limitations. Teachers should select item formats based on a variety of factors, such as the objectives, specific skill to be measured, and the ability level of the students. Many objectives are better measured with certain item formats.

Decisions about what scoring procedure or procedures to use are somewhat dependent on the choice of item formats. Student responses to some item formats must be hand-scored, whether they are recorded directly on the test itself or on a separate answer sheet or in a booklet. The teacher should decide whether the time and resources available for scoring a test suggest that hand-scoring or machine-scoring would be preferable.

The best way to ensure measurement validity of a teacher-constructed test is to develop a test blueprint, also known as a *test plan* or a *table of specifications*, before building the test itself. The elements of a test blueprint include (a) a list of the major topics or instructional objectives that the test will cover, (b) the level of complexity of the task to be assessed, and (c) the emphasis each topic will have, indicated by number or percentage of items or points. The test blueprint serves several important functions. It is a useful tool for guiding the work of the item writer so that sufficient items are developed at the appropriate level to test important content areas and objectives. The blueprint also should be used to inform students about the nature of the test and how they should prepare for it.

After developing the test blueprint, the teacher writes the test items that correspond to it. Regardless of the selected item formats, the teacher should follow some general rules that contribute to the development of high-quality test items. Those rules were discussed in the chapter.

Because teacher-made tests typically measure students' maximum performance rather than their typical performance, teachers should create conditions under which students will be able to demonstrate their best possible performance. These conditions include adequate preparation of the students to take the test. Adequate preparation includes information, skills, and attitudes that will facilitate students' maximum performance on the test.

■ References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York, NY: Longman.
- Bosher, S. D. (2009). Removing language as a barrier to success on multiple-choice exams. In S. D. Bosher & M. D. Pharris (Eds.), *Transforming nursing education: The culturally inclusive environment* (pp. 259–284). New York, NY: Springer Publishing Company.
- Bosher, S., & Bowles, M. (2008). The effects of linguistic modification on ESL students' comprehension of nursing course test items. *Nursing Education Perspectives*, 29, 165–172.
- Brookhart, S. M., & Nitko, A. J. (2019). *Educational assessment of students* (8th ed.). New York, NY: Pearson Education.
- Miller, M. D., Linn, R. L., & Gronlund, N. E. (2013). *Measurement and assessment in teaching* (11th ed.). Upper Saddle River, NJ: Prentice Hall.
- National Council of State Boards of Nursing. (2019). *2019 NCLEX® test plan*. Chicago, IL: Author. Retrieved from https://www.ncsbn.org/2019_RN_TestPlan-English.pdf
- Poorman, S. G., Mastorovich, M. L., Molcan, K. L., & Liberto, T. L. (2011). *Good thinking: Test taking and study skills for nursing students* (3rd ed.). Pittsburgh, PA: STAT Nursing Consultants.
- Waltz, C. F., Strickland, O. L., & Lenz, E. R. (2010). *Measurement in nursing and health research* (4th ed.). New York, NY: Springer Publishing Company.

TRUE–FALSE AND MATCHING

There are different ways of classifying types of test items. One way is based on how they are scored—objectively or subjectively. An example of an objectively scored item is multiple choice: There is one correct or best answer. By choosing that answer, students receive a particular score, such as one point. Essay items are subjectively scored: The teacher judges the quality of the response based on criteria or a rubric for scoring. Another way is to group test items by the type of response required of the test-taker. Selected-response items require the test-taker to select the correct or best answer from options provided by the teacher. Examples of these items include true–false, matching exercises, multiple choice, and multiple response. Constructed-response items, such as completion and essay, ask the test-taker to supply an answer rather than choose from options already provided. For each of the item formats presented in this book, a number of principles should be considered when writing them. Although important principles are described, the lists presented are not intended to be inclusive; other sources on test construction might include additional helpful suggestions for writing test items. Appendix A provides a quick reference guide to writing varied types of test items with examples.

In addition to test items, other assessment strategies are written assignments, cases, presentations, and projects that student complete. These strategies and others, including methods for evaluating clinical performance, are discussed in the book. Tests and other types of assessment strategies may be used at the beginning of a course to determine whether students have the prerequisite knowledge for achieving the outcomes or whether they have already met them. With courses that are competency based, students can then progress to the next area of instruction. Tests, quizzes, and other assessment strategies also are used during the instruction to provide the basis for formative assessment. This form of assessment is to monitor learning progress, provide feedback to students, fill in gaps in their learning, and suggest additional learning activities as needed. The goal of formative assessment is to support students on their learning: It is assessment *for* learning (Brookhart & Nitko, 2019). At the end of the course, tests and other assessment strategies determine whether students have achieved the outcomes and are the basis for assigning a grade in the course.

■ True–False

A true–false item consists of a statement that the student needs to judge as either true or false. In some items, students also correct false statements or supply a rationale as to why the statement is true or false. True–false items are most effective for recall of facts and specific information but also may be used to test the student’s understanding of the information. They are not intended for assessing complex thinking, which Miller, Linn, and Gronlund (2013) suggested is the main limitation of using these items. Each true–false item represents a declarative sentence stating a fact or principle and asks the learner to decide whether it is true or false, right or wrong, correct or incorrect. Some authors refer to this type of test item as *alternate response*, allowing for these varied response formats. For affective outcomes, agree–disagree might be used, asking the learner to agree or disagree with a value-based statement.

There are different opinions as to the value of true–false items. Although some authors express concern over the low level of testing, focusing on recall of facts and the opportunity for guessing, others indicate that true–false items provide an efficient means of examining student acquisition of knowledge in a course. With true–false items, students can respond to a large number of items in a short time. For that reason, true–false items are useful to include on a quiz or test, and they also provide a way of testing a wide range of content. These items are easy to write and to score.

Although true–false items are relatively easy to construct, the teacher should avoid using them to test meaningless information. Designed to examine student recall and understanding of *important* facts and principles, true–false items should not be used to evaluate memorization of irrelevant information. Prior to constructing these items, the teacher should ask: Is the content assessed by the true–false item important when considering the course outcomes? Does the content represent knowledge taught in the class or through other methods of instruction? Do the students need an understanding of the content to progress through the course and for their further learning?

The main limitation to true–false items is guessing. Because one of the two responses has to be correct, the probability that a student will answer the item correctly is 50%. However, the issue with guessing is not as much of a problem as it seems. With no knowledge of the facts being tested, on a 10-point quiz, the student would only be expected to answer five of the items or 50% correctly. Brookhart and Nitko (2019) suggested that few students in a course respond to test items with blind or completely random guessing. Most students have some knowledge of the subject even if they need to guess an answer. It also is difficult to obtain an adequate score on a test by using random guessing only. Although students have a 50/50 chance of guessing a correct answer on one true–false item, the probability of guessing correctly on a test with many items is small. For example, if a test has 10 true–false items, a student who guesses blindly on all of those items has less than six chances out of 100 of having 80% or more of the items correct.

Writing True–False Items

The following discussion includes some important principles for the teacher to consider when constructing true–false items.

1. *The true–false item should test recall of important facts and information.* The teacher should avoid constructing items that test trivia and meaningless information. The content should be worth knowing and be important in relation to the course outcomes.
2. *The statement should be true or false without qualification—unconditionally true or false.* The teacher should be able to defend the answer without conditions.
3. *Avoid words such as usually, sometimes, often, and similar terms.* Because these words typically are used more often in true statements, they give students clues to the correct response. Along the same lines, avoid words such as *never, always, all, and none*, which often signal a false response.
4. *Avoid terms that indicate an infinite degree or amount such as large.* They can be interpreted differently by students.
5. *Each item should include one idea to be tested rather than multiple ones.* When there are different propositions to be tested, each should be designed as a single true–false item.
6. *Items should be worded precisely and clearly.* The teacher should avoid long statements with different qualifiers and focus the sentence instead on the main idea to be tested. Long statements take time for reading and do not contribute to testing student knowledge of an important fact or principle. Brookhart and Nitko (2019) recommended that short statements work best for true–false items because the item can be focused on essential content to be assessed.
7. *Avoid the use of negatives, particularly in false statements.* They are confusing to read and may interfere with student ability to understand the statement. For instance, the item “It is not normal for a 2-year-old to demonstrate hand preference” (true) would be stated more clearly as, “It is normal for a 2-year-old to demonstrate hand preference” (false). If negative words such as *not* and *no* must be used in the item, they should be highlighted.
8. *With a series of true–false items, statements should be similar in length.* The teacher may be inclined to write longer true sentences than false ones in an attempt to state the concept clearly and precisely.
9. *Check that the answers to true–false items are not ordered in a noticeable pattern on the test.* For example, the teacher should avoid arranging the items in a pattern in which the answers would be TFTF or FTTFTT.

10. *Decide how to score true–false items prior to administering them to students.* In some variations of true–false items, students correct false statements; for this type, the teacher should award 2 points, 1 for identifying the statement as false and 1 for correcting it. With items of this type, the teacher should not reveal the point value of each item because this would cue students that 2-point items are false.

Sample items follow.

For each of the following statements, select T if the statement is true and F if the statement is false:

- T F Type 1 diabetes was formerly called insulin-dependent diabetes. (T)
- T F Hypothyroidism is manifested by lethargy and fatigue. (T)
- T F The most common congenital heart defect in children is tetralogy of Fallot. (F)

Variations of True–False Items

There are many variations of true–false items that may be used for testing. One variation is to ask the students to correct false statements. Students may identify the words that make a statement false and insert words to make it true. In changing the false statement to a true one, students may write in their own corrections or choose words from a list supplied by the teacher. One other modification of true–false items is to have students include a rationale for their responses, regardless of the statement being true or false. This provides a means of testing their understanding of the content.

For all of these variations, the directions should be clear and specific. Some examples follow.

1. If the statement is true, select T and do no more. If the statement is false, select F and underline the word or phrase that makes it false.
 T F Tetany occurs with increased levels of calcium.
 Because this statement is false, the student should select F and underline the word “increased”:
 T F Tetany occurs with increased levels of calcium. (F)
2. If the statement is true, select T and do no more. If the statement is false, select F, underline the word or phrase that makes it false, and write in the blank the word or phrase that would make it true.
 T F Canned soups are high in potassium.

T F Fresh fruits and vegetables are low in sodium.

In the first example, because the statement is false, the student should select F, underline “potassium,” and write “sodium” in the blank to make the statement true. In the second example, because the statement is true, the student should only select T:

T F Canned soups are high in potassium. (F)

Sodium

T F Fresh fruits and vegetables are low in sodium. (T)

3. If the statement is true, select T and do no more. If the statement is false, select F and select the *correct* answer from the list that follows the item.

T F Bradycardia is a heart rate less than 80 beats per minute.
40, 50, 60, 100

Because the statement is false, the student should select both F and 60:

T F Bradycardia is a heart rate less than 80 beats per minute. (F)
40, 50, 60, 100

4. If the statement is true, select T and explain why it is true. If the statement is false, select F and explain why it is false.

T F One purpose of Kegel exercises is to strengthen the pubococcygeal muscles. (T)

One other variation of true-false items is called the *multiple true-false* item. This is a cross between a multiple-choice and a true-false item. Multiple true-false items have an incomplete statement followed by several phrases that complete it; learners indicate which of the phrases form true or false statements. This type of item clusters true-false statements under one stem. However, rather than selecting one answer as in a multiple-choice item, students decide whether each alternative is true or false (Brookhart & Nitko, 2019). Directions for answering these items should be clear, and the phrases should be numbered consecutively because they represent individual true-false items. As with any true-false item, the phrases that complete the statement should be unequivocally true or false.

Sample items follow.

The incomplete statements are followed by several phrases. Each of the phrases completes the statement and makes it true or false. If the completed statement is true, select T. If the completed statement is false, select F.

A patient with a below-the-knee amputation should:

- T F 1. Avoid walking until fitted with a prosthesis. (F)
- T F 2. Keep the residual limb elevated at all times. (F)
- T F 3. Exercise the arms against resistance. (T)
- T F 4. Not sit for prolonged periods of time. (T)

The taxonomy of the cognitive domain includes the:

- T F 5. Applying level. (T)
- T F 6. Remembering level. (T)
- T F 7. Calculating level. (F)
- T F 8. Acting level. (F)
- T F 9. Analyzing level. (T)
- T F 10. Manipulating level. (F)
- T F 11. Creating level. (T)

■ Matching Exercises

Matching exercises consist of two parallel columns in which students match terms, phrases, sentences, or numbers from one column to the other. In a matching exercise, students identify the one-to-one correspondence between the columns. One column includes a list of premises (for which the match is sought); the other column (from which the selection is made) is a list of responses (Brookhart & Nitko, 2019). The basis for matching responses to premises should be stated explicitly in the directions with the exercise. The student identifies pairs based on the principle specified in these directions. With some matching exercises, differences between the premises and responses are not apparent, such as matching a list of laboratory studies with their normal ranges, and the columns could be interchanged. In other exercises, however, the premises include descriptive phrases or sentences to which the student matches shorter responses.

Matching exercises lend themselves to testing categories, classifications, groupings, definitions, and other related facts. They are most appropriate for measuring facts based on simple associations (Miller et al., 2013). One advantage of a matching exercise is its ability to test a number of facts that can be grouped together rather

than designing a series of individual items. For instance, the teacher can develop one matching exercise on medications and related side effects rather than a series of individual items on each medication. A disadvantage of matching exercises, however, is the focus on remembering facts and specific information, although in some courses this reflects an important outcome of learning.

Writing Matching Exercises

Matching exercises are intended for categories, classifications, and information that can be grouped in some way. An effective matching exercise requires the use of homogeneous content with responses that are plausible for the premises. Responses that are not plausible for some premises provide clues to the correct match. Principles for writing matching exercises include:

1. *Develop a matching exercise around homogeneous content.* All of the premises and responses to be matched should relate to that content, for example, all laboratory tests and values, all terms and definitions, and all types of drugs and drug actions. This is the most important principle in writing a matching exercise (Miller et al., 2013).
2. *Include an unequal number of premises and responses to avoid giving a clue to the final match.* If there are the same number of premise statements and responses, and students can match all but one remaining pair, they automatically get that last match correct. It is typical for there to be more responses than premises, but the number of responses may be limited by the maximum number of answer options allowed on a scannable answer sheet or online test. In that case, the teacher may need to write more premises than responses.
3. *Use short lists of premises and responses.* This makes it easier for the teacher to identify items from the same content area, and it saves students reading time. With a long list of items to be matched, it is difficult to review the choices and pair them with the premises. It also prohibits recording the answers on a scannable form and may affect constructing these items for an online test. Miller et al. (2013) recommended using four to seven items in each column. A longer list might be used for some exercises, but no more than 10 items should be included in either column (p. 188).
4. *For matching exercises with a large number of responses, the teacher should develop two separate matching exercises.* Otherwise, students spend too much time reading through the options.
5. *Directions for the matching exercises should be clear and state explicitly the basis for matching the premises and responses.* This is an important principle in developing these items. Even if the basis for matching seems self-evident, the directions should include the rationale for matching the columns.

6. *Directions should specify whether each response may be used once, more than once, or not at all.* Matching items can be developed in which students match one response to one premise, with at least one “extra” response remaining to avoid giving a clue to the final match. Items also can be written in which students can use the responses more than once or not at all. The directions should be unambiguous about the selection of responses.
7. *Place the longer premises on the left and shorter responses on the right.* This enables the students to read the longer statement first, then search on the right for the correct response, which often is a single word or a few words.
8. *The order in which the premises and responses are listed should be alphabetical, numerical, or follow some other logical order.* Listing them in a logical order eliminates clues and saves students time when reading through the item. If the lists have another logical order, however, such as dates and sequential steps of a procedure, then they should be organized in that order. Numbers, quantities, and similar types of items should be arranged in decreasing or increasing order.
9. *The entire matching exercise should be placed on the same page and in the same box if online.* This prevents students from missing possible responses that are on the next page or screen.

Sample matching items are found in Exhibits 4.1 and 4.2.

EXHIBIT 4.1

SAMPLE MATCHING ITEM

Directions: For each definition in Column A, select the proper term in Column B. Use each letter only once or not at all.

COLUMN A (PREMISES)	COLUMN B (RESPONSES)
<u> b </u> 1. Attaching a particular response to a specific stimulus	a. Cognitive styles
<u> f </u> 2. Believing that one can respond effectively in a situation	b. Conditioning
<u> g </u> 3. Changing gradually behavioral patterns	c. Empowerment
<u> d </u> 4. Observing a behavior and its consequences and attempting to behave similarly	d. Modeling
<u> a </u> 5. Varying ways in which individuals process information	e. Self-care
	f. Self-efficacy
	g. Shaping

EXHIBIT 4.2**SAMPLE MATCHING ITEM**

Directions: For each type of insulin in Column A, identify its peak action in Column B. Responses in Column B may be used once, more than once, or not at all.

COLUMN A	COLUMN B
<u> c </u> 1. Regular	a. Long acting
<u> b </u> 2. NPH	b. Intermediate acting
<u> a </u> 3. Glargine	c. Short acting
<u> a </u> 4. Detemir	

■ Summary

This chapter described how to construct two types of test items: true-false and matching exercises, including their variations. A true-false item consists of a statement that the student judges as either true or false. In some forms, students correct a false statement or supply a rationale as to why the statement is true or false. True-false items are most effective for recall of facts and specific information, but also may be used to test the student's understanding of an important principle or concept.

Matching exercises consist of two parallel columns in which students match terms, phrases, sentences, or numbers from one column to the other. One column includes a list of premises and the other column, from which the selection is made, contains the responses. The student identifies pairs based on the principle specified in the directions. Matching exercises lend themselves to testing categories, classifications, groupings, definitions, and other related facts. As with true-false items, they are most appropriate for testing recall of specific information.

■ References

- Brookhart, S. M., & Nitko, A. J. (2019). *Educational assessment of students* (8th ed.). New York, NY: Pearson Education.
- Miller, M. D., Linn, R. L., & Gronlund, N. E. (2013). *Measurement and assessment in teaching* (11th ed.). Upper Saddle River, NJ: Pearson Education.

MULTIPLE-CHOICE AND MULTIPLE-RESPONSE

This chapter focuses on two other kinds of selected-response items: multiple-choice and multiple-response. Multiple-choice items, which have one correct or one best answer, are used widely in nursing and in other fields. This test-item format includes a question or incomplete statement, followed by a list of options that answer the question or complete the statement. Multiple-response items are designed similarly, although more than one answer may be correct. Both of these test-item formats may be used for assessing learning at the remembering, understanding, applying, and analyzing levels, making them adaptable for a wide range of content and learning outcomes. Appendix A provides a quick reference guide to writing multiple-choice and multiple-response items and some examples of these items.

■ Multiple-Choice Items

Multiple-choice items can be used for assessing many types of outcomes. Some of these include

- Knowledge of facts, specific information, and principles
- Definitions of terms
- Understanding of content and concepts
- Application of concepts to patient scenarios
- Analysis of data and clinical situations
- Comparison and selection of varied treatments and interventions
- Judgments and decisions about actions to take in clinical and other situations

Multiple-choice items are particularly useful in nursing to assess outcomes that require students to apply knowledge and analyze data and situations. With multiple-choice items, the teacher can introduce *new* information requiring application of concepts and theories or analytical thinking to respond to the questions. Experience

with multiple-choice testing provides essential practice for students who will later encounter this type of item on licensure, certification, and other commercially prepared examinations. Multiple-choice items also allow the teacher to sample the course content more easily than with items such as essay questions, which require more time for responding. In addition, multiple-choice tests can be electronically scored and analyzed.

Although there are many advantages to multiple-choice testing, there are also disadvantages. First, these items are difficult to construct, particularly at the higher cognitive levels. Developing items to test memorization of facts is much easier than designing ones to measure use of knowledge in a new situation and skill in analysis. As such, many multiple-choice items are written at the lower cognitive levels, focusing only on remembering and understanding. Second, teachers often have difficulty developing plausible distractors. These distractors—also spelled *distractions*—are the incorrect alternatives that seem plausible for test-takers who have not adequately learned the content. If a distractor is not plausible, it provides an unintended clue to the test-taker that it is not the correct response. Third, it is often difficult to identify only one correct answer. For these reasons, multiple-choice items are time-consuming to construct.

Some critics of multiple-choice testing suggest that essay and similar types of questions to which students develop a response provide a truer measure of learning than items in which students choose from available options. However, multiple-choice items written at the applying and analyzing levels require *use* of knowledge and analytical thinking to make a selection from the available options. For items at higher cognitive levels, test-takers need to compare options and make a judgment about the correct or best response. Pham et al. (2018) found that multiple-choice items can test higher level learning as well as short-answer items.

Writing Multiple-Choice Items

There are three parts to a multiple-choice item, each with its own set of principles for development: (a) stem, (b) answer, and (c) distractors. Table 5.1 indicates each of these parts.

The stem is the lead-in phrase in the form of a question or an incomplete statement that relies on the alternatives for completion. Following the stem is a list of alternatives or options for the learner to consider and choose from. These alternatives are of two types: the answer, also called the *keyed response*, which is the correct or best response to answer the question or complete the statement, and distractors, which are the incorrect alternatives. The purpose of the distractors, as the word implies, is to *distract* students who are unsure of the correct answer. Suggestions for writing

TABLE 5.1 Parts of a Multiple-Choice Item

AN EARLY AND COMMON SIGN OF PREGNANCY IS	STEM IN FORM OF INCOMPLETE STATEMENT
OPTIONS OR ALTERNATIVES	
a. Amenorrhea b. Morning sickness c. Spotting d. Tenderness of the breasts	Answer Distractor Distractor Distractor
IN WHICH OF THE FOLLOWING GROUPS DOES RAYNAUD'S DISEASE OCCUR MOST FREQUENTLY?	STEM IN FORM OF QUESTION
OPTIONS OR ALTERNATIVES	
a. Men between 20 and 40 y old b. Men between 50 and 70 y old c. Women between 20 and 40 y old d. Women between 50 and 70 y old	Distractor Distractor Answer Distractor

each of these parts are considered separately because they have different principles for construction.

Stem

The stem is the question or incomplete statement to which the alternatives relate. Whether the stem is written in question form or as an incomplete statement, the most important quality is its clarity. The test-taker should be able to read the stem and know what to look for in the alternatives without having to read through them. Thus, after reading the stem, the learner should understand the intent of the item and what type of response the teacher expects (Brookhart & Nitko, 2019). One other important consideration in writing the stem is to ensure that it presents a problem or situation that relates to the learning outcome being assessed. Guidelines for writing the stem are as follows:

1. *The stem should present clearly and explicitly the problem to be solved.* The student should not have to read the alternatives to understand the question or the intent of the incomplete statement. The stem should provide sufficient information for answering the question or completing the statement. An example of this principle follows:

Cataracts:

- a. are painful.
- b. may accompany coronary artery disease.
- c. occur with aging.¹
- d. result in tunnel vision.

The stem of this question does not present the problem associated with cataracts that the alternatives address. As such, it does not guide the learner in reviewing the alternatives. In addition, the options are dissimilar, which is possible because of the lack of clarity in the stem; alternatives should be similar. Also, the keyed response implies that aging is the only cause of cataracts, which is not accurate and may confuse students who know the content well. One possible revision of this stem is:

The causes of cataracts include:

- a. aging.¹
- b. arteriosclerosis.
- c. hemorrhage.
- d. iritis.

After writing the item, the teacher should cover the alternatives and read the stem alone. Does it explain the problem and direct the learner to the alternatives? Is it complete? Could it stand alone as a short-answer item? In writing the stem, always include the nature of the response, such as, “Which of the following *interventions, signs and symptoms, treatments, data*,” and so forth. A stem that simply asks, “Which of the following?” does not provide clear instructions as to what to look for in the options.

2. *Although the stem should be clear and explicit, it should not contain extraneous information unless the item is developed for the purpose of identifying significant versus insignificant data.* Otherwise, the stem should be brief, including only necessary information. Long stems that include irrelevant information take additional time for reading. This point can be illustrated as follows, using the previous cataract item:

¹Correct answer.

You are caring for an elderly man who lives alone but has frequent visits from his daughter. He has congestive heart failure and some shortness of breath. Your patient was told recently that he has cataracts. The causes of cataracts include:

- a. aging.¹
- b. arteriosclerosis.
- c. hemorrhage.
- d. iritis.

In this stem, the background information about the patient is irrelevant to the problem addressed. If subsequent items were to be written about the patient's other problems, related nursing interventions, the home setting, and so forth, then this background information might be presented as a scenario in a context-dependent item set (see Chapter 7, Assessment of Higher Level Learning).

Stems also should not be humorous; laughing during the test can distract students who are concentrating. If one of the distractors is humorous, it will be recognized as implausible and eliminated as an option, increasing the chance of guessing the correct answer from among the remaining alternatives.

3. *Avoid inserting information in the stem for instructional purposes.* In the example that follows, the definition of cataract has no relevance to the content tested, that is, the causes of cataracts. The goal of testing is to evaluate outcomes of learning, not to teach new information, as in this example:

Cataracts are an opacity of the lens or capsule of the eye leading to blurred and eventual loss of vision. The causes of cataracts include:

- a. aging.¹
- b. arteriosclerosis.
- c. hemorrhage.
- d. iritis.

4. *If words need to be repeated in each alternative to complete the statement, shift them to the stem.* This is illustrated as follows:

An early and common sign of pregnancy:

- a. *is amenorrhea.*¹
- b. *is morning sickness.*
- c. *is spotting.*
- d. *is tenderness of the breasts.*

The word *is* may be moved to the stem:

An early and common sign of pregnancy *is*:

- a. amenorrhea.¹
- b. morning sickness.
- c. spotting.
- d. tenderness of the breasts.

Similarly, a word or phrase repeated in each alternative does not test students' knowledge of it and should be included in the stem. An example follows:

Clinical manifestations of Parkinson's disease include:

- a. decreased perspiration, tremors at rest, and *muscle rigidity*.¹
- b. increased salivation, *muscle rigidity*, and diplopia.
- c. *muscle rigidity*, decreased salivation, and nystagmus.
- d. tremors during activity, *muscle rigidity*, and increased perspiration.

This item does not test knowledge of muscle rigidity occurring with Parkinson's disease because it is included with each alternative. The stem could be revised as follows:

Clinical manifestations of Parkinson's disease include *muscle rigidity* and which of the following signs and symptoms?

- a. Decreased salivation and nystagmus.
 - b. Increased salivation and diplopia.
 - c. Tremors at rest and decreased perspiration.¹
 - d. Tremors during activity and increased perspiration.
5. *Do not include key words in the stem that would clue the student to the correct answer.* This point may be demonstrated in the earlier question on cataracts.

You are caring for an *elderly* patient who was told recently that he has cataracts. The causes of cataracts include:

- a. *aging*.¹
- b. arteriosclerosis.
- c. hemorrhage.
- d. iritis.

In this item, informing the student that the patient is elderly provides a clue to the correct response.

6. *Avoid the use of negatively stated stems, including words such as no, not, and except.* Negatively stated stems are sometimes unclear; in addition, they require a change in thought pattern from selections that represent correct and best responses to ones reflecting incorrect and least likely responses. Most stems may be stated positively, asking for the correct or best response rather than the exception. If there is no acceptable alternative to a negatively stated stem, consider rewriting the item in a different format, such as true–false, completion, or multiple response. If a negatively phrased item must be used, the negative word should be only in the stem or only in an alternative, but not in both of them, and the negative word should be underlined or placed in CAPITAL LETTERS (Brookhart & Nitko, 2019).
7. *The stem and alternatives that follow should be consistent grammatically.* If the stem is an incomplete statement, each option should complete it grammatically; if not, clues may be provided as to the correct or incorrect responses. It is also important to check carefully that a consistent verb form is used with the alternatives. An example follows:

Your patient is undergoing a right carotid endarterectomy. Prior to surgery, which information would be most important to collect as a baseline for the early recovery period? Her ability to:

- a. follow movements with her eyes.
- b. move all four extremities.¹
- c. rotating her head from side to side.
- d. swallow and gag.

Option “c” provides a grammatical clue by not completing the statement “Her ability to.” The item may be revised easily:

Your patient is undergoing a right carotid endarterectomy. Prior to surgery, which information would be most important to collect as a baseline for the early recovery period? Her ability to:

- a. follow movements with her eyes.
- b. move all four extremities.¹
- c. rotate her head from side to side.
- d. swallow and gag.

8. *Avoid ending stems with “a” or “an” because these often provide grammatical clues as to the option to select.* It is usually easy to rephrase the stem to eliminate the “a” or “an.” For instance,

Narrowing of the aortic valve in children occurs with *an*:

- a. aortic stenosis.¹
- b. atrial septal defect.
- c. coarctation of the aorta.
- d. patent ductus arteriosus.

Ending this stem with “an” eliminates alternatives “c” and “d” because of an obvious lack of grammatical agreement. The stem could be rewritten by deleting the “an”:

Narrowing of the aortic valve in children occurs with:

- a. aortic stenosis.¹
- b. atrial septal defect.
- c. coarctation of the aorta.
- d. patent ductus arteriosus.

Ending the stem with “a or an” or “a/an” is not a satisfactory alternative because these formats require test-takers to reread each alternative with “a” first and then “an,” thereby increasing reading time unnecessarily.

9. *If the stem is a statement completed by the alternatives, begin each alternative with a lowercase letter and place a period after it because it forms a sentence with the stem. At the end of the stem, use a comma or colon as appropriate. Use uppercase letters to begin alternatives that do not form a sentence with the stem. If the stem is a question, place a question mark at the end of the stem.*
10. *Each multiple-choice item should be independent of the others. The answer to one item should not be dependent on a correct response to another item, and the test-taker should not have to read another item to correctly interpret the item at hand. In the following example, the meaning of the second-item stem cannot be understood without referring to the stem of the first item:*
 1. You are the community health nurse developing a teaching plan for a 45-year-old man who was treated in the emergency department for an asthma attack. Which action should be implemented *FIRST*?
 - a. Assess other related health problems.
 - b. Determine his level of understanding of asthma.¹
 - c. Review with him treatments for his asthma.
 - d. Teach him actions of his medications.

2. On your second home visit, the patient is short of breath. Which of these statements indicates a need for further instruction?
 - a. “I checked my peak flow because I’m not feeling good.”
 - b. “I have been turning on the air conditioner at times like this.”
 - c. “I tried my Advair because my chest was feeling heavy.”¹
 - d. “I used my nebulizer mist treatment for my wheezing.”

A better format would be to develop a series of multiple-choice items that relate to a patient scenario, clinical situation, or common data set (context-dependent item set), with directions that indicate the items that pertain to the given context. This item format is discussed in Chapter 7, Assessment of Higher Level Learning.

11. *Write the stem so that the alternatives are placed at the end of the incomplete statement.* An incomplete statement with a blank in the middle, which the options then complete, interrupts the reading and may be confusing for the students to read and follow (Brookhart & Nitko, 2019). For example:

The nurse should check the _____ for a patient receiving warfarin.

- a. activated clotting time
- b. complete blood cell count
- c. international normalized ratio¹
- d. partial thromboplastin time

This item would be easier to read for students if the alternatives were placed at the end of the statement:

For a patient receiving warfarin, the nurse should check the:

- a. activated clotting time.
- b. complete blood cell count.
- c. international normalized ratio.¹
- d. partial thromboplastin time.

Alternatives

Following the stem in a multiple-choice item is a list of alternatives or options, which include (a) the correct or best answer and (b) distractors. There are varying recommendations in the literature as to the number of alternatives to include,

ranging from three to five. An early meta-analysis and other studies have found that three options (one correct answer and two distractors) were usually sufficient: Some distractors were selected by so few students that they did not function as distractors (Edwards, Arthur, & Bruce, 2012; Kilgour & Tayyaba, 2016; Rodriguez, 2005; Tarrant, Ware, & Mohammed, 2009).

Raymond, Stevens, and Bucak (2019) examined the use of the three-option format on the physician licensure examination. They analyzed 3,360 distractors across 840 items on the examination. Their findings confirmed earlier studies that most multiple-choice items have one or more distractors that function poorly (students do not choose them) and omitting the least functional distractor did not affect test reliability. The findings provided minimal support for five options but did not clearly point to whether multiple-choice items for high-stakes tests in healthcare are best developed using three (one correct answer and two distractors) or four (one correct answer and three distractors) options (Raymond et al., 2019). They proposed continued use of the four-option format, consistent with the examples in this book.

Four options allow for one correct or best answer and three plausible distractors. Many standardized tests use four alternatives. In writing multiple-choice items, however, if one of the distractors is not plausible, it is better to use a three-option item (Dehnad, Hayedeh, & Hosseini, 2014; Raymond et al., 2019; Rodriguez, 2005; Tarrant & Ware, 2012). General principles for writing the alternatives follow:

1. *The alternatives should be similar in length, detail, and complexity.* It is important to check the number of words included in each option for consistency in length. Frequently the correct answer is the longest because the teacher attempts to write it clearly and specifically. Brookhart and Nitko (2019) suggested that the testwise student may realize that the longest or most complete response is the correct answer without having the requisite knowledge to make this choice. In that case, the teacher should either shorten the correct response or add similar qualifiers to the distractors so that they are similar in length as well as in detail and complexity.

Although there is no established number of words by which the alternatives may differ from each other without providing clues, one strategy is to count the words in each option and attempt to vary them by no more than a few words. This will ensure that the options are consistent in length. In the sample item, the correct answer is longer than the distractors, which might provide a clue for selecting it.

You are assessing a 14-year-old girl who appears emaciated. Her mother describes the following changes: resistance to eating and 20-lb weight loss over the past 6 weeks. It is most likely that the patient resists eating for which of the following reasons?

- a. Complains of recurring nausea
- b. Describes herself as “fat all over” and fearful of gaining weight¹
- c. Has other gastrointestinal problems
- d. Seeks her mother’s attention

The correct answer can be shortened to: Is fearful of gaining weight.

2. *In addition to consistency in length, detail, and complexity, the options should have the same number of parts.* The answer in the previous question is not only longer than the other options but also includes two parts, providing another clue. In the example that follows, including two causes in option “a” provides a clue to the answer. Revising that option to only “aging” avoids this.

Causes of cataracts include:

- a. aging and steroid therapy.¹
- b. arteriosclerosis.
- c. hemorrhage.
- d. iritis.

3. *The alternatives should be consistent grammatically.* The answer and distractors should be similar in structure and terminology. Without this consistency in format, the test-taker may be clued to the correct response or know to eliminate some of the options without being familiar with the content. In the following sample item, the student may be clued to the correct answer “a” because it differs grammatically from the others:

You are making a home visit with a new mother who is breastfeeding. She tells you that her nipples are cracked and painful. Which of the following instructions should be given to the mother?

- a. Put the entire areola in the baby’s mouth during feeding.¹
- b. The baby should be fed less frequently until the nipples are healed.
- c. There is less chance of cracking if the nipples are washed daily with soap.
- d. Wiping off the lotion on the nipples before feeding the baby may help.

4. *The alternatives should sample the same domain, for instance, all symptoms, all diagnostic tests, all nursing interventions, varying treatments, and so forth.* A study by Ascalon, Meyers, Davis, and Smits (2007) examined the effects on

item difficulty of different ways of writing the item stem and homogeneity of the alternatives. They found no differences in item difficulty when the stem was written as a statement versus a question. However, when alternatives of a multiple-choice item were similar, it increased the item difficulty. It is likely that when responses are dissimilar from the correct response, learners can easily eliminate them as options. In the example that follows, option “b” is not a nursing diagnosis, which may clue the student to omit it as a possibility.

You are working in the emergency department, and your patient is having difficulty breathing. His respiratory rate is 40, heart rate 140, and oxygen saturation 90%. He also complains of a headache. Which of the following nursing diagnoses is of greatest priority?

- a. Activity intolerance
 - b. Chronic obstructive pulmonary disease
 - c. Impaired gas exchange¹
 - d. Pain
5. *Avoid including opposite responses among the options.* This is often a clue to choose between the opposites and not consider the others. A sample item follows:

The nurse should determine the correct placement of a nasogastric tube by:

- a. asking the patient to swallow.
- b. aspirating gastric fluid from the tube.
- c. confirming the placement with an x-ray.¹
- d. confirming the placement with capnography.

In this example, the correct response is opposite one of the distractors, which clues the student to select one of these alternatives. In addition, options “c” and “d” begin with “confirming,” which may provide a visual clue to choose between them. To avoid this possible clue, the first distractor in the example could be reworded to form a second pair of opposites:

The nurse should determine the correct placement of a nasogastric tube by:

- a. aspirating air from the tube.
 - b. aspirating gastric fluid from the tube.
 - c. confirming the placement with an x-ray.¹
 - d. confirming the placement with capnography.
6. *Arrange the options in a logical or meaningful order.* The order can be alphabetical, numerical, or chronological (Brookhart & Nitko, 2019). Arranging

the options in this way tends to randomly distribute the position of the correct response rather than the answer occurring most often in the same location, for example, “b” or “c,” throughout the test. It also helps students locate the correct response more easily when they have an answer in mind.

7. *Options with numbers, quantities, and other numerical values should be listed sequentially, either increasing or decreasing in value, and the values should not overlap.* When alternatives overlap, a portion of a distractor may be correct, resulting in more than one correct answer. These problems are apparent in the sample item that follows:

The normal range for potassium in adults is:

- a. 2.5–4.5 mEq/L.
- b. .5–3.5 mEq/L.
- c. 3.5–5.1 mEq/L.¹
- d. 1.5–4.5 mEq/L.

The values in these options overlap, and the alternatives would be easier to review if they were arranged sequentially from decreasing to increasing values. Laboratory and other values should be labeled appropriately, such as hemoglobin 14.0 g/dL. A revision of the prior item follows:

The normal range for potassium in adults is:

- a. .5–1.5 mEq/L.
- b. 2.0–3.2 mEq/L.
- c. 3.5–5.1 mEq/L.¹
- d. 5.5–7.5 mEq/L.

8. *Each option should be placed on a separate line for ease of student reading.* If answers are recorded on a separate answer sheet, the teacher should review the format of the sheet ahead of time so that responses are identified as “a” through “d” or 1 through 4 as appropriate. Usually items are numbered and responses are lettered to prevent clerical errors when students use a separate answer sheet.
9. *Use the option of “call for assistance” and “notify the physician” sparingly.* It is not known how they act as distractors in multiple-choice items. In addition, some teacher-made tests may overuse this option as the correct answer, conditioning students to select it without considering the other alternatives.

Correct or Best Answer

In a multiple-choice item, there is one answer to be selected from among the alternatives. In some instances, the best rather than the correct answer should be chosen. Considering

that judgments are needed to arrive at decisions about patient care, items can ask for the best or most appropriate response from those listed. Best answers are valuable for more complex and higher level learning such as with items written at the application and analysis levels. Even though best-answer items require a judgment to select the best option, there can be only one answer, and there should be consistency in the literature and among experts as to that response. A colleague can review the items, without knowing the answers in advance, to ensure that they are correct.

Suggestions follow for writing the correct or best answer. These suggestions are guided by the principle that the students should not be able to identify the correct response and eliminate distractors because of the way the stem or alternatives are written.

1. *Review the alternatives carefully to ensure that there is only one correct response.* For example:

Symptoms of increased intracranial pressure include:

- a. blurred vision*
- b. decreased blood pressure.
- c. disorientation.¹
- d. increased pulse.

In this sample item, both “a” and “c” are correct; a possible revision follows:

Symptoms of increased intracranial pressure include:

- a. blurred vision and decreased blood pressure.
- b. decreased blood pressure and increased pulse.
- c. disorientation and blurred vision.¹
- d. increased pulse and disorientation.

2. *Review terminology included in the stem carefully to avoid giving a clue to the correct answer.* Key words in the stem, if also used in the correct response, may clue the student to select it. In the following example, “sudden weight loss” is in both the stem and the answer:

An elderly patient with *sudden weight loss*, thirst, and confusion is seen in the clinic. Which of the following signs would be indicative of dehydration?

- a. Below normal temperature
- b. Decreased urine specific gravity
- c. Increased blood pressure
- d. *Sudden weight loss*¹

The question could be revised by omitting “sudden weight loss” in the stem.

An elderly patient with dry skin, thirst, and confusion is seen in the clinic. Which of the following signs would also be indicative of dehydration?

- a. Below normal temperature
 - b. Decreased urine specific gravity
 - c. Increased blood pressure
 - d. Sudden weight loss¹
3. *The correct answer should be randomly assigned to a position among the alternatives to avoid favoring a particular response choice.* Some teachers may inadvertently assign the correct answer to the same option (e.g., “c”) or, over a series of items, a pattern may develop from the placement of the correct answers (e.g., “a, b, c, d, a, b, c, d”). As indicated earlier in the discussion of how to write the options, this potential clue can be avoided by listing the alternatives in a logical or meaningful order such as alphabetical, numerical, or chronological. However, the teacher also should double check the position of the correct answers on a test to confirm that they are more or less randomly distributed.
 4. *The answers should not reflect the opinion of the teacher but instead should be the ones with which experts agree or are the most probable responses.* The answers should be consistent with the literature and not be answers chosen arbitrarily by the teacher. Alternatively, a specific authority may be referenced in the stem (e.g., “According to the Centers for Disease Control and Prevention”).

Distractors

Distractors are the incorrect but plausible options offered. Distractors should appeal to learners who lack the knowledge to respond to the question without confusing those who know the content. If the option is obviously wrong, then there is no reason to include it as an alternative. Because the intent of the distractors is to appeal to learners who have not mastered the content, at least some of the students should choose each option or the distractors should be revised for the next administration of the test.

Each alternative should be appropriate for completing the stem. Hastily written distractors may be clearly incorrect, may differ in substance and format from the others, and may be inappropriate for the stem, providing clues as to how to respond. They also may result in a test item that does not measure the students’ learning.

When writing a multiple-choice item, it is sometimes difficult to identify enough plausible distractors to have the same number of options for each item on the test. However, rather than using a filler that is obviously incorrect or would not be seriously considered by the students, the teacher should use fewer options on that item. The goal is to develop plausible alternatives, ones that attract at least some of the

students, rather than filler alternatives that no one chooses. Thus, for some items, there may be only three alternatives, even though the majority of questions on that test use four. The goal, however, is to develop three plausible distractors so that most items have at least four responses from which to choose.

In writing distractors, it is helpful to think about common errors that students make, phrases that “sound correct,” misperceptions students have about the content, and familiar responses not appropriate for the specific problem in the stem. Another way of developing distractors is to identify, before writing any of the options, the content area or domain to which all the responses must belong, for example, all nursing interventions. If the stem asks about nursing measures for a patient with acute pneumonia, the distractors might be interventions for a patient with asthma that would not be appropriate for someone with pneumonia.

Terms used in the stem also give ideas for developing distractors. For example, if the stem asks about measures to avoid increasing anxiety in a patient who is delusional, the distractors may be interventions for a delusional patient that might inadvertently increase or have no effect on anxiety, or interventions useful for decreasing anxiety but not appropriate for a patient with a delusional disorder. Another strategy for developing distractors is to identify the category to which all alternative responses must belong. For a stem that asks about side effects of erythromycin, plausible distractors may be drawn from side effects of antibiotics as a group. Suggestions for writing distractors include:

1. *The distractors should be consistent grammatically and should be similar in length, detail, and complexity with each other and the correct answer.* Examples were provided earlier in the chapter. The distractors should be written with the same specificity as the correct response. If the correct response is “quadratus plantae,” distractors that are more general, such as “motor,” may be a clue not to choose that option.
2. *The distractors should sample the same content area as the correct answer.* When types of options vary, they may clue the student as to the correct response or to eliminate a particular distractor. In the following example, options “a,” “b,” and “c” pertain to factors in the workplace. Because option “d” relates to diet, it may clue the student to omit it. A better alternative for “d” would be another factor to assess in the work setting such as how tiring the job is.

In planning teaching for a patient with a hiatal hernia, which of these factors should be assessed?

- a. Amount of lifting done at work¹
- b. Number of breaks allowed
- c. Stress of the job
- d. Use of high-sodium foods

3. *Avoid using “all of the above” and “none of the above” in a multiple-choice item.* As distractors, these contrast with the direction of selecting one correct or best response. With “all of the above” as a distractor, students aware of one incorrect response are clued to eliminate “all of the above” as an option. Similarly, knowledge of one correct alternative clues students to omit “none of the above” as an option. Often teachers resort to “all of the above” when they are unable to develop a fourth option, although it is better to rephrase the stem or use three options as discussed earlier.

The “none of the above” option is appropriate for multiple-choice items for which students perform calculations. By using “none of the above,” the teacher avoids giving clues to students when their incorrect answer is not listed with the options. In the following example, the student would need to know the correct answer to identify that it is not among the alternatives:

You are working in a pediatrician’s office, and a mother calls and asks you how many drops of acetaminophen to give to her infant. The order is for 40 mg every 12 hours, but the container she has at home is 80 mg/0.8 mL. You should tell the mother to give:

- a. 1 dropperful
 - b. 1 teaspoon
 - c. 1.5 mL in a 3-mL syringe
 - d. None of the above¹
4. *Omit terms such as “always,” “never,” “sometimes,” “occasionally,” and similar words from the distractors.* These general terms often provide clues as to the correctness of the option. Terms such as *always* and *never* suggest that the alternatives are incorrect because rarely does a situation occur always or never, particularly in patient care.
5. *Avoid using distractors that are essentially the same.* In the following example, alternatives “a” and “c” are essentially the same. If “rest” is eliminated as an option, the students are clued to omit both of these. In addition, the correct response in this item is more general than the others and is not specific to this particular student’s health problems.

A student comes to see the school nurse complaining of a severe headache and stiff neck. Which of the following actions would be most appropriate?

- a. Ask the student to rest in the clinic for a few hours.
- b. Collect additional data before deciding on interventions.¹
- c. Have a family member take the student home to rest.
- d. Prepare to take the student to urgent care.

The item could be revised as follows:

A student comes to see the school nurse complaining of a severe headache and stiff neck. Which of the following actions would be most appropriate?

- a. Ask the student to rest in the clinic for a few hours.
- b. Check for numbness or loss of strength in the student's arms.¹
- c. Prepare to take the student to urgent care.
- d. Send the student back to class after medicating for pain.

Variation of Multiple-Choice Items

A multiple-choice item can be combined with short-answer or essay. In this format, after answering a multiple-choice item, students develop a rationale for why their answer is correct and the distractors are incorrect. The teacher should award 1 point for correctly identifying the answer and other points for providing an acceptable rationale. For example:

Your patient is ordered 30 mg of Roxanol (morphine sulfate 20 mg/mL) every 4 hours for severe pain. Which of the following actions should be taken?

- a. Call the physician about the dose.
- b. Dilute in 500 mL normal saline.
- c. Give the morphine as ordered.¹
- d. Hold if the respiratory rate is less than 10.

In the following space, provide a rationale for why your answer is the best one and why the other options are not appropriate.

■ Multiple Response

In these item formats several alternatives may be correct, and students choose either all of the correct alternatives or the best combination of alternatives. Multiple-response items are included on the NCLEX® (National Council Licensure Examination) as one type of item format (National Council of State Boards of Nursing, 2019). On the NCLEX and other types of computerized tests, students select all of the options that apply by checking the box that precedes each option, as in the following example:

The preliminary diagnosis for your patient, a 20-year-old college student, is meningitis. Which signs and symptoms should you anticipate finding? Select all that apply.

- ☐ 1. Abdominal tenderness
- ☒ 2. Fever
- ☐ 3. Lack of pain with sudden head movements
- ☒ 4. Nausea and vomiting
- ☒ 5. Nuchal rigidity
- ☒ 6. Sensitivity to light
- ☐ 7. Sudden bruising in neck area

The principles for writing multiple-response items are the same as for writing multiple-choice items. Suggestions for writing these items include the following:

1. *The combination of alternatives should be plausible.* Options should be logically combined rather than grouped randomly.
2. *The alternatives should be used a similar number of times in the combinations.* If one of the alternatives is in every combination, it is obviously correct; this information should be added to the stem as described earlier in the chapter. Similarly, limited use of an option may provide a clue to the correct combination of responses. After grouping responses, each letter should be counted to be sure that it is used a similar number of times across combinations of responses and that no letter is included in every combination.
3. *The responses should be listed in a logical order, for instance, alphabetically or sequentially, for ease in reviewing.* Alternatives are easier to review if shorter combinations are listed before longer ones.

A sample item follows:

Which actions are appropriate to prevent a retained surgical item?

1. If an instrument can be disassembled, all parts are counted.
 2. The circulating nurse counts sponges that are added to the case.
 3. The surgical technologist orders an x-ray if the needle count is incorrect.
 4. Used sponges are displayed in a clear plastic holder.
- a. 1, 2
 - b. 1, 4¹
 - c. 2, 3
 - d. 3, 4
 - e. 1, 3, 4

■ Summary

This chapter described the development of multiple-choice and multiple-response items. Multiple-choice items, with one correct or best answer, are used widely in nursing and other fields. This test-item format includes a question or incomplete statement followed by a list of options that answer the question or complete the statement. Multiple-response items are designed similarly although more than one answer may be correct, or students may be asked to select one answer with the best combination of alternatives. All of these item formats may be used for evaluating learning at the remembering, understanding, applying, and analyzing levels, making them adaptable for a wide range of content and outcomes.

Multiple-choice items are important for testing the application of nursing knowledge in simulated clinical situations and clinical judgment. Because of their versatility, they may be integrated easily within most testing situations.

There are three parts in a multiple-choice item, each with its own set of principles for development: (a) stem, (b) answer, and (c) distractors. The stem is the lead-in phrase presented in the form of a question or an incomplete statement that relies on the alternatives for completion. Following the stem is a list of alternatives, which are options for the learner to consider and choose from. These alternatives are of two types: the answer, which is the correct or best option to answer the question or complete the statement, and distractors, which are the incorrect yet plausible alternatives. Suggestions for writing each of these parts were presented in the chapter and were accompanied by sample items. Other examples and a quick reference guide are found in Appendix A.

The ability to write multiple-choice items is an important skill for the teacher to develop. This is a situation in which “practice makes perfect.” After writing an item, the teacher should have colleagues read it and make suggestions for revision. Although time-consuming to develop, multiple-choice items are an important means for assessing learning in nursing.

■ References

- Ascalon, M. E., Meyers, L. S., Davis, B. W., & Smits, N. (2007). Distractor similarity and item-stem structure: Effects on item difficulty. *Applied Measurement in Education*, 20, 153–170. doi:10.1080/08957340701301272
- Brookhart, S. M., & Nitko, A. J. (2019). *Educational assessment of students* (8th ed.). New York, NY: Pearson Education.
- Dehnad, A., Hayedeh, N., & Hosseini, A. F. (2014). A comparison between three- and four-option multiple choice questions. *Procedia-Social and Behavioral Sciences*, 98, 398–403. doi:10.1016/j.sbspro.2014.03.432
- Edwards, B. D., Arthur, W., & Bruce, L. L. (2012). The 3-option format for knowledge and ability multiple-choice tests: A case for why it should be more commonly used in personnel testing. *International Journal of Selection and Assessment*, 20, 65–81. doi:10.1111/j.1468-2389.2012.00580.x

- Kilgour, J. M., & Tayyaba, S. (2016). An investigation into the optimal number of distractors in single-best answer exams. *Advances in Health Sciences Education*, 21, 571–585. doi:10.1007/s10459-015-9652-7
- National Council of State Boards of Nursing. (2019). *2019 NCLEX-RN® detailed test plan*. Chicago, IL: Author.
- Pham, H., Trigg, M., Wu, S., O'Connell, A., Harry, C., Barnard, J., & Devitt, P. (2018). Choosing medical assessments: Does the multiple-choice question make the grade? *Education for Health*, 31, 65–71. doi:10.4103/efh.EfH_229_17
- Raymond, M. R., Stevens, C., & Bucak, S. D. (2019). The optimal number of options for multiple-choice questions on high-stakes tests: Application of a revised index for detecting nonfunctional distractors. *Advances in Health Sciences Education*, 24(1), 141–150. doi:10.1007/s10459-018-9855-9
- Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, 24(2), 3–13.
- Tarrant, M., & Ware, J. (2012). A framework for improving the quality of multiple-choice assessment. *Nurse Educator*, 37, 98–104. doi:10.1097/NNE.0b013e31825041d0
- Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and nonfunctioning distractors in multiple-choice questions: A descriptive analysis. *BMC Medical Education*, 9, 40–47. doi:10.1186/1472-6920-9-40

SHORT-ANSWER (FILL-IN-THE-BLANK) AND ESSAY

Short-answer and essay questions are examples of constructed-response items. With these items, the test-taker supplies an answer rather than selecting from options already provided. Because students supply the answers, this type of item reduces the chance of guessing.

Short-answer items can be answered with a word, phrase, or number. There are two types of short-answer items: question and completion. One format presents a question that students answer in a few words or phrases. With the other format, completion or fill in the blank, students are given an incomplete sentence that they complete by inserting a word or words in the blank space. In an essay item, the student develops a more extended response to a question or statement. Essay tests and written assignments use writing as the means of expressing ideas, although with essay items the focus of assessment is the content of the answer rather than the writing ability. Short-answer and essay items are described in this chapter. A quick reference guide to writing these items is provided in Appendix A.

■ Short Answer

Short-answer items can be answered by a word, phrase, or number. The two types of short-answer items—question and completion—also referred to as fill-in-the-blank, are essentially the same except for format.

With the question format, students answer a question in a few words or phrases. Calculations may be included for the teacher to review the process that the student used to arrive at an answer. The questions may stand alone and have no relationship to one another, or comprise a series of questions in a similar content area.

Completion items consist of a statement with a key word or words missing; students fill in the blank to complete it. Other types of completion items ask students to perform a calculation and record the answer, or to order a list of responses. Completion items are appropriate for recall of facts and specific information and for calculations. To complete the statement, the student recalls missing facts, such as a word or short phrase, or records the solution to a calculation problem. Although

completion items appear easy to construct, they should be designed in such a way that only one answer is possible. If students provide other correct answers, the teacher needs to accept them.

Fill-in-the-blank calculation and ordered-response items are two of the alternate item formats used on the NCLEX® (National Council Licensure Examination). Fill-in-the-blank items ask candidates to perform a calculation and type in the answer. All answers are scored as right or wrong. With ordered-response items, candidates answer a question by rank ordering options or placing a list of responses in the proper order (National Council of State Boards of Nursing, 2019). For example, students might be given a list of characteristics of pressure ulcers or the steps of a procedure and asked to put them in the order in which they occur. On a computerized test, such as the NCLEX, candidates can click an option (unordered side of the screen) and drag and drop it in the correct order (on the ordered response side of the screen), or highlight an option and use the arrow keys to arrange the options in the correct order. However, this same format can be used on a paper-and-pencil test with students writing the order on the answer sheet.

Short-answer items are useful for measuring student ability to interpret data, use formulas correctly, complete calculations, and solve mathematical-type problems. Brookhart and Nitko (2019) described another type of short-answer format, association variety, which provides a list of terms, diagram, or photograph for which students recall relevant labels, numbers, or symbols. For example, students might be given a list of medical terms and asked to list their abbreviations.

Writing Short-Answer Items

Suggestions for developing short-answer items are as follows:

1. *Questions and statements should not be taken verbatim from textbooks, other readings, and handouts students received.* These materials may be used as a basis for designing short-answer items, but taking exact wording from them may result in testing only recall of meaningless facts out of context. Such items measure memorization of content and may or may not be accompanied by the student's comprehension of it.
2. *Phrase the item so that a unique word, series of words, or number must be supplied to complete it.* Only one correct answer should be possible to complete the statement.
3. *Write questions that are specific and can be answered in a few words, phrases, or short sentences.* The question, "What is insulin?" does not provide sufficient direction as to how to respond; asking instead, "What is the peak action time of NPH insulin?" results in a more specific answer.
4. *Before writing the item, think of the correct answer first and then write a question or statement for that answer.* Although the goal is to develop an item with only one correct response, students may identify other correct answers.

For this reason, it is necessary to develop a scoring sheet with all possible correct answers, and rescore student responses as needed if students provide additional correct answers that the teacher did not anticipate.

5. *Fill-in-the-blank items requiring calculations and solving mathematical-type problems should include in the statement the type of answer and degree of specificity desired.* For example, the question may ask the test-taker to convert pounds to kilograms, rounding the answer to one decimal point.
6. *For a statement with a key word or words missing, place the blank at or near the end of the statement.* This makes it easier for students to complete. It is also important to watch for grammatical clues in the statement, such as “a” versus “an” and singular versus plural, prior to the blank, which might give clues to the intended response. If more than one blank is included in the statement, they should be of equal lengths. Use of more than two blanks should be avoided because there may be insufficient information remaining to permit students to grasp the nature of the task they are to perform.
7. *When students need to write longer answers, provide for sufficient space.* In some situations, longer responses might indicate that the item is actually an essay item, and the teacher then should follow principles for constructing and scoring essay items.
8. *Even though a blank space is placed at the end of the statement, the teacher may direct the student to record one-word answers in blanks to the left of the items, thereby facilitating scoring.* For example,

_____ 1. *Streptococcus pneumoniae* and *Staphylococcus aureus* are examples of _____ bacteria.

Following are some examples of question and completion (or fill-in-the-blank) formats of short-answer items:

What congenital cardiac defect results in communication between the pulmonary artery and the aorta? _____

Two types of metered-dose inhalers used for the treatment of bronchial asthma are:

Six Sigma is a quality-improvement model used in healthcare. List two other models.

1. _____

2. _____

You are caring for a patient who weighs 128 lb. She is ordered 20 mcg/kg of an intravenous (IV) medication. What is the correct dose in micrograms?

Answer: _____

■ Essay Item

In an essay test, students construct responses to items based on their understanding of the content. With this type of test item, varied answers may be possible depending on the concepts selected by the student for discussion and the way in which they are presented. Essay items provide an opportunity for students to select content to discuss, present ideas in their own words, and develop an original and creative response to an item. Essay items are useful for assessing complex learning outcomes and higher levels of learning. Higher level responses, however, are more difficult to evaluate and score than answers reflecting recall of facts.

Although some essay items are developed to assess recall of facts and specific information, they are more appropriate for higher levels of learning. Miller, Linn, and Gronlund (2013) recommended that essay items should be used primarily for learning outcomes that cannot be measured adequately through selected-response items. Essay items are effective for evaluating students' ability to apply concepts, analyze situations and ideas, and develop creative solutions to problems using multiple sources of information. Higher level thinking skills and more complex learning outcomes such as analysis, synthesis, and evaluation are better assessed with essays, because students need to select the information they want to include and formulate their own responses rather than choosing one from a list already provided (Gierl, Latifi, Lai, Boulais, & De Champlain, 2014; Rios & Wang, 2018).

Although essay items use writing as the medium for expression, the intent is to assess student understanding of specific content rather than to judge writing ability in and of itself. Written assignments are better suited to evaluating the ability of students to write effectively; these are described in Chapter 9. Low-level essay items are similar to short-answer items and require precise responses. An example of a low-level essay item is: "Describe three signs of increased intracranial pressure in children younger than 2 years old." Broader and higher level essay items, however, do not limit responses in this way and differ clearly from short-answer items, such as "Defend the statement 'parents should be required to follow childhood vaccine recommendations.'" Essay items may be written to assess a wide range of outcomes. These include:

- Comparing, such as comparing the side effects of different medications
- Outlining steps to take and protocols to follow
- Explaining in one's own words a situation or statement

- Discussing topics
- Applying concepts and evidence to a clinical scenario and explaining their impact on potential decisions
- Analyzing patient data and clinical situations
- Critiquing different interventions based on current evidence
- Developing plans and proposals drawing on multiple sources of information
- Analyzing nursing and healthcare trends
- Arriving at decisions about issues and actions to take, with a rationale added
- Analyzing ethical issues, possible decisions, and their consequences
- Developing arguments for and against a particular position or decision

As with other types of test items, the outcome to be assessed provides the framework for developing the essay item. From the outcome, the teacher develops a clear and specific item to elicit information about student achievement. If the outcome focuses on application of concepts to clinical practice, then the essay item should examine the ability to apply knowledge to a clinical situation. The item should be stated clearly so that the students know what they should write about. If it is ambiguous, students may perceive the need to write all they know about a topic.

Issues With Essay Tests

Although essay items are valuable for examining the ability to select, organize, and present ideas and they provide an opportunity for creativity and originality in responding, they are limited by low reliability and other issues associated with their scoring. The teacher should have an understanding of these issues because they may influence the decision to use essay items. Strategies are provided later in the chapter for addressing some of these issues.

Limited Ability to Sample Content

By their nature, essay items do not provide an efficient means of sampling course content as compared with objective items. Often, only a few essay items can be included on a test, considering the time it takes for students to formulate their thoughts and prepare an open-ended response, particularly when the items are intended for assessing higher levels of learning. As a result, it is difficult to assess all of the different content areas in a nursing course using essay items.

When the outcomes are memorization and remembering facts, essay items should not be used because there are more efficient means of measuring such outcomes. Instead, good essay items ask students to use higher level thinking skills and organize and express ideas (Brookhart & Nitko, 2019). Essay items are best used for responses requiring originality.

Unreliability in Scoring

The major limitation of essay items is the lack of consistency in evaluating responses. Scoring answers is a complex process, and studies have shown that essay responses are scored differently by different teachers and even the same teacher may not be consistent with scoring a response at a different time (Miller et al., 2013; Rios & Wang, 2018). Some teachers are more lenient or critical than others regardless of the criteria established for scoring. Even with preset criteria, teachers may evaluate answers differently, and scores may vary when the same teacher reads the paper again. Miller et al. (2013) suggested that frequently the reasons for unreliability in scoring are the failure of the faculty member to identify the specific outcomes being assessed with the essay item and lack of a well-defined rubric for scoring (p. 238).

Factors such as misspelled words and incorrect grammar may affect scoring beyond the criteria to which they relate. In scoring the student's response, it is important to focus on the substantive content and not be influenced by how the response is written. Brookhart and Nitko (2019) recommended that writing style, spelling, and grammar should be scored separately to avoid blending this evaluation with a judgment about the student's knowledge of the subject. These areas can be specified as a separate score on the rubric used for the evaluation and be given less weight than the substantive content of the answer.

The unreliability with scoring depends on the type of essay item. When the essay item is highly focused and structured, such as "Describe three side effects of bronchodilators," there is greater reliability in scoring. These lower level items also could be classified as short answer. Less restrictive essay items allowing for freedom and creativity in responding have lower rater reliability than more restricted ones. Items asking students to analyze, defend, judge, evaluate, and create products are less reliable in terms of scoring the response. There are steps the teacher can take, though, to improve reliability, such as defining the content to be included in a "correct" answer and using a scoring rubric. These are presented later in the chapter.

Carryover Effects

Another issue in evaluating essay items is a carryover effect in which the teacher develops an impression of the quality of the answer from one item and carries it over to the next response. If the student answers one item well, the teacher may be influenced to score subsequent responses at a similarly high level; the same situation may occur with a poor response. For this reason, it is best to read all students' responses to one item before evaluating the next one. Miller et al. (2013) suggested that reading all the answers to one item at a time improves scoring accuracy by keeping the teacher focused on the standards of each item. It also avoids carrying over an impression of the quality of the student's answer to one item onto the scoring of the next response.

The same problem can occur with written assignments. The teacher's impression of the student can carry over from one paper to the next. When scoring essay tests and grading papers, the teacher should not know whose paper it is.

Halo Effect

There may be a tendency in evaluating essay items to be influenced by a general impression of the student or feelings about the student, either positive or negative, that create a halo effect when judging the quality of the answers. For instance, the teacher may hold favorable opinions about the student from class or clinical practice and believe that this learner has made significant improvement in the course, which in turn might influence the scoring of responses. For this reason, essay tests should be scored anonymously by asking students to identify themselves by an assigned or selected number rather than by their names. Names can be matched with numbers after scoring is completed.

Rater Drift

Essay tests read early in a scoring session may be scored higher than those read near the end because of teacher fatigue and time constraints. Brookhart and Nitko (2019) described the problem of rater drift, the tendency of the teacher to gradually stray from the scoring criteria. Over time the teacher may not pay attention to the specific criteria or may apply them differently to each response. In scoring essay items, the teacher needs to check that the rubric and standards for grading are implemented equally for each student. Teachers should read papers in random order and read each response twice before computing a score. After scoring the responses to a question, the teacher should rearrange the papers to avoid being influenced by their order. It also is important to stop periodically and confirm that the responses read later are scored consistently with those read early (Brookhart & Nitko, 2019). Another potential issue with scoring is the teacher may tend to award scores near the mean or drift between scoring too severely or leniently regardless of the quality of the response (Rios & Wang, 2018). An awareness by the teacher of potential issues with scoring essay items can lead to better practices with assessing student answers.

Time

One other issue in using essay items is the time it takes for students to answer them and for teachers to score them. In writing essay items, the teacher should estimate how long it will take to answer each item, erring on allowing too much time rather than too little. Students should be told approximately how long to spend on each item so they can pace themselves (Miller et al., 2013).

Scoring essay items also can be a pressing issue for teachers, particularly if the teacher is responsible for large numbers of students. Considering that responses should be read twice, the teacher should consider the time required for scoring responses when planning for essay tests. Scoring software is available that can scan an essay and score the response; however, this software is not designed for assessing constructed responses in specialized fields such as nursing.

Student Choice of Items

Some teachers allow students to choose a subset of essay items to answer, often because of limited time for testing and to provide options for students. For example, the teacher may include four items on the care of patients with type 1 diabetes and ask students to answer two of them. However, Miller et al. (2013) cautioned against this practice because when students choose different items to answer, they are actually taking different tests. The option to choose items to answer also may affect measurement validity.

Restricted-Response Essay Items

There are two types of essay items: restricted response and extended response. Although the notion of freedom of response is inherent in essay items, there are varying degrees of freedom in responding to the items. At one end of the continuum is the restricted-response item, in which a few sentences are required for an answer. These are short-answer essays. At the other end is the extended-response item, in which students have freedom to express their own ideas and organize them as they choose. Responses to essay items typically fall between these two extremes.

In a restricted-response item, the teacher limits the student's answer by indicating the content to be discussed, the extent of discussion allowed, or both. For example, a specific patient problem might be identified and students asked questions about that problem, or the directions to the item might limit the response to one paragraph or page. With this type of essay item, the way in which the student responds is structured by the teacher. In writing restricted-response items, the teacher might include specific material with the item, such as patient data, a description of a clinical scenario, a summary of evidence, a description of issues associated with clinical practice, and extracts from the literature, to cite a few. Students read, analyze, and interpret this accompanying material, then answer questions about it.

Examples of restricted-response items follow:

- Define *patient-centered care*. Limit your definition to one paragraph.
- Select one environmental health problem and describe its potential effects on the community. Do not use an example presented in class. Limit your discussion to one page.

- Compare metabolic and respiratory acidosis. Include the following in your response: definitions, precipitating factors, clinical manifestations, diagnostic tests, and interventions.
- Your patient is 76 years old and 1-day postoperative following a femoral popliteal bypass surgery. Name two complications the patient could experience at this time and discuss why they are potential problems. List two nursing interventions for this patient to prevent these complications during the early recovery period with related evidence.
- Describe three pathological changes that characterize chronic obstructive pulmonary disease.

Extended-Response Essay Items

Extended-response essay items are less restrictive and as such provide an opportunity for students to decide how to respond: They can organize ideas in their own ways, arrive at judgments about the content, and demonstrate the ability to communicate ideas effectively in writing. With these types of items, the teacher may assess students' ability to develop their own ideas and express them creatively, integrate learning from multiple sources in responding, and evaluate the ideas of others based on predetermined criteria. Because responses are not restricted by the teacher, assessment is more difficult. This difficulty, however, is balanced by the opportunity for students to express their own ideas and for the teacher to assess higher level learning. As such, extended-response essay items provide a means of assessing more complex learning not possible with selected-response items. The teacher may decide to allow students to respond to these items outside of class. Sample items include:

- Critique arguments for and against the practice of suspending do-not-resuscitate status when patients undergo surgery. Based on your critique, state which position you believe is the strongest and provide a rationale supporting your choice.
- The fall rate on your unit has increased in the past 3 months. Develop a plan for analyzing this occurrence with a rationale to support your action plan.
- You receive a call in the allergy clinic from a mother who describes her son's problems as "having stomach pains" and "acting out in school." She asks you whether these problems may be due to his allergies. How would you respond to this mother? How would you manage this call? Include a rationale for your response with evidence to support your decisions.
- Describe an integrated model of palliative care and why this is important for children living with a life-threatening or terminal condition. What principles serve as a foundation for an integrated model of palliative care? List and discuss each principle and why it is important in pediatric palliative care.

Writing Essay Items

Essay items should be reserved for outcomes that cannot be assessed effectively through multiple-choice and other selected-response formats. With essays, students can demonstrate their higher level thinking and ability to integrate varied sources of information and concepts. Suggestions for writing essay items are as follows.

1. *Develop essay items that require synthesis of the content.* Avoid items that students can answer by merely summarizing the readings and class or online discussions without thinking about the content and applying it to new situations. Assessing students' recall of facts and specific information may be accomplished more easily using other formats, such as true–false and matching, rather than essay.
2. *Phrase items clearly.* The item should direct learners in their responses and should not be ambiguous. Exhibit 6.1 provides sample stems for essay items based on varied types of learning outcomes. Framing the item to make it as specific as possible is accomplished more easily with restricted-response items. With extended-response items, the teacher may provide directions as to the type of response intended without limiting the student's own thinking about the answer. In the example that follows, there is minimal guidance as to how to respond; the revised version, however, directs students more clearly as to the intended response without limiting their freedom of expression and originality.

Example: Evaluate an article describing a nursing research study.

Revised version: Select an article reporting the results of a nursing research study. Critique the study, specifying the criteria you used to evaluate it. Based on your evaluation, describe whether the research provides evidence for nursing practice and rate the quality of the evidence. Include a rationale supporting your answer, including the system used for rating the evidence.

3. *Prepare students for essay tests.* This can be accomplished by asking students thought-provoking questions; engaging students in critical discussions about the content; and teaching students how to apply concepts to clinical situations, to compare approaches, and to arrive at judgments about patients and issues. Practice in synthesizing content from different sources, presenting ideas logically, and using creativity in responding to situations will help students prepare to respond to essay items in a testing situation. This practice may come through discussions in class, in clinical practice, and online; written assignments; and small-group activities. For students lacking experience with essay tests, the teacher may use sample items for formative purposes, providing feedback to students about the adequacy of their responses.

EXHIBIT 6.1**SAMPLE STEMS FOR ESSAY ITEMS***Comparing*

Compare the side effects of . . . methods for . . . interventions for
 Describe similarities and differences between
 What do . . . have in common?
 Group these medications . . . signs and symptoms

Outlining Steps

Describe the process for . . . procedure for . . . protocol to follow for
 List steps in order for

Explaining and Summarizing

Explain the importance of . . . relevance of
 Identify and discuss
 Explain the patient's responses within the framework of
 Provide a rationale for
 Discuss the most significant points of
 Summarize the relevant data, evidence,
 What are the major causes of . . . reasons for . . . problems associated with
 Describe the potential effects of . . . possible responses to . . . problems that might result from

Applying Concepts to a Situation

Analyze the situation using . . . theory/framework.
 Using the theory of . . . , explain the patient's/family's responses.
 Identify and discuss . . . using relevant concepts.
 Describe a clinical situation that demonstrates the concept of

Analyzing

Discuss the significance of
 Identify relevant data with supporting rationale.
 Identify and describe additional data needed for decision-making.
 Describe possible patient problems with rationale.
 What hypotheses may be formed?
 Compare nursing interventions based on evidence.
 Describe multiple nursing interventions for this patient with supporting rationale.
 Provide a rationale for . . . , evidence for
 Critique the nurse's responses to this patient.
 Describe errors in assumptions made about . . . errors in reasoning
 Analyze the situation and describe possible alternate actions.
 Identify all possible decisions, consequences of each, your decision, and supporting rationale.

Developing Plans and Proposals

Develop a plan for . . . discharge plan
 Develop a proposal for . . . protocol for

(continued)

EXHIBIT 6.1**SAMPLE STEMS FOR ESSAY ITEMS (continued)**

Based on concepts of . . . , develop a plan for . . . proposal for

Develop a new approach for

Design multiple interventions for

Analyzing Issues

Identify a significant issue in healthcare and describe implications for nursing practice.

Analyze this issue and implications for

In light of these trends, what changes would you propose?

Critique the nurse's/physician's/healthcare team's/patient's decisions in this situation. What other approaches are possible? Why?

Analyze the ethical issue facing the nurse. Compare multiple decisions possible and consequences of each. Describe the decision you would make and why.

Identify issues for this patient/family/community and strategies for resolving them.

Stating Positions

What would you do and why?

Identify your position about . . . and defend it.

Develop an argument for . . . and against

Develop a rationale for

Do you support this position? Why or why not?

Do you agree or disagree with . . . ? Include a rationale.

Specify the alternative actions possible. Which of these alternatives would be most appropriate and why? What would you do and why?

4. *Tell students about apportioning their time to allow sufficient time for answering each essay item.* In writing a series of essay items, consider carefully the time needed for students to answer them and inform students of the estimated time before they begin the examination. In this way, students may gauge their time appropriately. Indicating the point value of each essay item also will guide students to use their time appropriately, spending more time on and writing longer responses to items that carry greater weight.
5. *Score essay items that deal with the analysis of issues according to the rationale that students develop rather than the position they take on the issue.* Students should provide a sound rationale for their position, and the evaluation should focus on the rationale rather than on the actual position.
6. *Avoid the use of optional items and student choice of items to answer.* As indicated previously, this results in different subsets of tests that may not be comparable.
7. *In the process of developing the item, write an ideal answer to it.* The teacher should do this while drafting the item to determine whether it is appropriate, clearly stated, and reasonable to answer in the allotted time frame. Save this ideal answer for use later in scoring students' responses.

8. *If possible, have a colleague review the item and explain how he or she would respond to it.* Colleagues can assess the clarity of the item and whether it will elicit the intended response.

Scoring Essay Items: Holistic Versus Analytic

There are two methods of scoring essay items: holistic and analytic. The holistic method involves reading the entire answer to each item and evaluating its overall quality. With the analytic method of scoring, the teacher separately scores individual components of the answer.

Holistic Scoring

With holistic scoring, the teacher assesses and scores the essay response as a whole without judging each part separately. There are different ways of scoring essays using the holistic method.

Relative scoring. One method of holistic scoring is to compare each student's answer with the responses of others in the group, using a relative standard. To score essay items using this system, the teacher quickly reads the answers to each item to gain a sense of how the students responded overall, then rereads the answers and scores them. Papers may be placed in groups reflecting degrees of quality with each group of papers receiving a particular score or grade.

Model answer. Another way is to develop a model answer for each item and then compare each student's response to that model. The model answer does not have to be written in narrative form, but can be an outline with the key points and elements that should be included in the answer. Before using a model answer for scoring responses, teachers should read a few papers to confirm that students' answers are consistent with what was intended.

Holistic scoring rubric. A third way of implementing holistic scoring is to use a scoring rubric, which is a guide for scoring essays, papers, written assignments, and other open-ended responses of students. Rubrics also can be used for grading posters, concept maps, presentations, and projects completed by students. The rubric consists of criteria used for evaluating the quality of the student's response. With holistic scoring, the rubric includes different levels of responses, with characteristics or descriptions of each level, and the related score. The student's answer is assigned the score associated with the one description within the rubric that best reflects its quality. The important concept in this method is that holistic scoring yields one overall score that considers the entire response to the item rather than scoring its component parts separately (Brookhart & Nitko, 2019).

Holistic rubrics are quicker to use for scoring because the teacher evaluates the overall response rather than each part of it. One disadvantage, though, is that they do

TABLE 6.1 Example of Holistic Scoring Rubric for Essay Item on Healthcare Issue

SCORE	DESCRIPTION
4	Presents thorough analysis of healthcare issue considering its complexities. Considers multiple perspectives in analysis. Analysis reflects use of theories and research. Discussion is well organized and supports analysis.
3	Analyzes healthcare issue. Considers different perspectives in analysis. Analysis reflects use of theories but not research. Discussion is organized and logical.
2	Describes healthcare issue but does not consider its complexities or different perspectives. Basic analysis of issue with limited use of theory. Discussion accurate but limited.
1	Does not clearly describe healthcare issue. No alternate perspectives considered. Limited analysis with no relevant theory or literature to support ideas. Errors in answer.
0	Does not identify the healthcare issue. No application of theory to understand issue. Errors in answer. Off-topic responses.

not provide students with specific feedback about their answers and can lead to more unreliability in scoring than with an analytic scoring rubric. An example of a holistic scoring rubric for an essay item is given in Table 6.1.

Analytic Scoring

In the analytic method of scoring, the teacher identifies the content that should be included in the answer and other characteristics of an ideal response. Each of these areas is assessed and scored separately. With analytic scoring, the teacher focuses on one characteristic of the response at a time (Miller et al., 2013). Often a detailed scoring plan is used that lists content to be included in the answer and other characteristics of the response to be judged. Students earn points based on how well they address each content area and the other characteristics, not their overall response.

Analytic scoring rubric. A scoring rubric also can be developed with points assigned for each of the content areas that should be included in the response and other characteristics to be evaluated. An analytic scoring rubric is useful in assessing essays and written work. First, it guides the teacher in judging the extent to which specified criteria have been met. Second, because this type of rubric specifies content to be included in a response with a related score, it can improve the accuracy of scoring essay answers and other written assignments. Third, a rubric creates a standardized method for grading assignments, which is valuable when there are different faculty members in a course, each grading his or her own students' work. Lastly, it provides feedback to students about the strengths and weaknesses of their response (Miller et al., 2013). An example of an analytic scoring rubric for the same essay item is found in Table 6.2.

TABLE 6.2 Example of Analytic Scoring Rubric for Essay Item on Healthcare Issue

SCORE	ANALYSIS OF ISSUE	MULTIPLE PERSPECTIVES	THEORY AND RESEARCH	PRESENTATION
4	Presents thorough analysis of healthcare issue considering its complexities.	Considers multiple perspectives in analysis.	Uses theories and research as basis for analysis.	Discussion is well organized and supports analysis.
3	Analyzes healthcare issue.	Considers a few varying perspectives.	Uses theories in analysis but no research.	Discussion is organized and logical.
2	Describes healthcare issue but does not consider its complexities.	Describes one perspective without considering other points of view.	Reports basic analysis of issue with limited use of theory.	Discussion is accurate but limited.
1	Does not clearly describe healthcare issue.	Considers no alternate perspectives.	Presents limited analysis with no relevant theories or literature to support ideas.	Discussion has errors in content.
0	Does not identify healthcare issue.	Considers no alternate perspectives.	Does not apply any theories in discussion.	Discussion has errors in content. May be off topic.
Score _____				

■ Criteria for Assessing Essay Items

The criteria for assessing essay items, regardless of the method, often address three areas: (a) content, (b) organization, and (c) process. Questions that guide assessment of each of these areas are:

- *Content:* Is relevant content included? Is it accurate? Are significant concepts and theories presented? Are hypotheses, conclusions, and decisions supported? Is the answer comprehensive?
- *Organization:* Is the answer well organized? Are the ideas presented clearly? Is there a logical sequence of ideas?
- *Process:* Was the process used to arrive at conclusions, actions, approaches, and decisions logical? Were different possibilities and implications considered? Was a sound rationale developed using relevant literature and theories?

Suggestions for Scoring

1. *Identify the method of scoring to be used prior to the testing situation and inform the students of it.*
2. *Specify in advance an ideal answer.* In constructing this ideal answer, review readings, the content presented in class and online, and other instructional activities completed by students. Identify content and characteristics required in the answer and assign points to them if using the analytic method of scoring.
3. *If using a scoring rubric, discuss it with the students ahead of time so that they are aware of how their essay responses will be judged.*
4. *Read a random sample of papers to get a sense of how the students approached the items and an idea of the overall quality of the answers. If students did not address an area in the scoring rubric, the teacher may revise the rubric and at a later time review the essay item for needed changes.*
5. *Score the answers to one item at a time.* For example, read and score all of the students' answers to the first item before proceeding to the second item. This procedure enables the teacher to compare responses to an item across students, resulting in more accurate and fairer scoring, and saves time by only needing to keep in mind one ideal answer at a time (Miller et al., 2013).
6. *Read each answer twice before scoring.* In the first reading, note omissions of major points from the ideal answer, errors in content, problems with organization, and problems with the process used for responding. Make notes about omissions, errors, and problems and record other comments on the students' paper in a format that can be modified if needed after reading the response a second time. When reading electronic versions of papers, these comments can be made using a word-processing application's review feature and easily changed after the second reading. After reading through all the answers to the question, begin the second reading for scoring purposes.
7. *Read papers in random order.*
8. *Use the same scoring system for all papers.*
9. *Read essay answers and other written assignments anonymously.* Develop a system for implementing this in the nursing education program, for instance, by asking the students to choose a code number.

10. *Cover the scores of the previous answers to avoid being biased about the student's ability.*
11. *For important decisions or if unsure about the scoring, have a colleague read and score the answers to improve reliability.* A sample of answers might be independently scored rather than the complete set of student tests.
12. *Adopt a policy on writing* (sentence structure, spelling, punctuation, grammar, neatness, and writing style in general) and determine whether the quality of the writing will be part of the score for the essay. Inform students of the policy in advance of the test. If writing is assessed, then it should be scored separately, and the teacher should be cautious not to let the writing style bias the evaluation of content and other characteristics of the response.

■ Summary

Short-answer items can be answered by a word, phrase, or number. There are two types of short-answer items: question and completion, also referred to as fill-in-the-blank. These items are appropriate for recall of facts and specific information. With short-answer items, students can be asked to interpret data, use formulas, complete calculations, and solve mathematical-type problems.

In an essay test, students construct responses to items based on their understanding of the content. With this type of test item, varied answers may be possible depending on the content selected by the student for the response and the way in which different areas of content are integrated. Essay items provide an opportunity for students to select content to discuss, to integrate content, to present ideas in their own words, and to develop original and creative responses to items. This freedom of response makes essay items particularly useful for complex and higher level outcomes.

There are two types of essay items: restricted response and extended response. In a restricted-response item, the teacher limits the student's answer by indicating the content to be discussed and frequently the amount of discussion allowed, for example, limiting the response to one paragraph or page. In an extended-response item, students have freedom of response, often requiring extensive writing. Although essay items use writing as the medium for expression, the intent is to assess student understanding of specific content rather than judge the writing ability in and of itself. Other types of assignments are better suited to assessing the ability of students to write effectively.

■ References

- Brookhart, S. M., & Nitko, A. J. (2019). *Educational assessment of students* (8th ed.). New York, NY: Pearson Education.
- Gierl, M. J., Latifi, S., Lai, H., Boulais, A. P., & De Champlain, A. (2014). Automated essay scoring and the future of educational assessment in medical education. *Medical Education*, 48, 950–962. doi:10.1111/medu.12517
- Miller, M. D., Linn, R. L., & Gronlund, N. E. (2013). *Measurement and assessment in teaching* (11th ed.). Upper Saddle River, NJ: Pearson Education.
- National Council of State Boards of Nursing. (2019). *2019 NCLEX-RN® test plan*. Chicago, IL: Author.
- Rios, J. A., & Wang, T. (2018). Essay items. In B. B. Frey (Ed.), *The SAGE encyclopedia of educational research, measurement, and evaluation*. Thousand Oaks, CA: Sage.

ASSESSMENT OF HIGHER LEVEL LEARNING

In preparing students to meet the needs of patients within the changing healthcare system, educators are faced with identifying essential content to teach in the nursing program. Mastery of this knowledge alone, however, is not enough. Students also need to develop cognitive skills for processing and analyzing information, comparing different approaches, weighing alternatives, and arriving at sound decisions. These cognitive skills include, among others, the ability to apply concepts to new situations, problem-solving, critical thinking, and clinical judgment. The purpose of this chapter is to present methods for assessing these higher levels of learning in nursing.

■ Higher Level Learning

One of the concepts presented in Chapter 1, Assessment and the Educational Process, was that outcomes can be organized in a cognitive hierarchy or taxonomy, with each level representing more complex learning than the previous one. Learning extends from simple remembering and understanding, which are lower level cognitive behaviors, to higher level thinking skills. Higher level cognitive skills include applying, analyzing, evaluating, and creating. With higher level thinking, students apply concepts and other forms of knowledge to new situations, use that knowledge to interpret patient needs and identify patient and other types of problems, and arrive at carefully thought-out judgments about actions to take.

The main principle in assessing higher level learning is to develop test items and other assessment methods that require students to apply knowledge and skills in a *new* situation (Brookhart & Nitko, 2019). Only then can the teacher assess whether the students are able to use what they have learned in a different context. Considering that patient characteristics, problems, and interventions often do not match the text-book descriptions, and health status can change quickly, students need to develop

their ability to think through clinical situations and arrive at the best possible decisions. By introducing novel materials into the assessment, the teacher can determine whether students have developed these cognitive skills.

■ Problem-Solving

In the practice setting, students are continually faced with patient and other clinical problems to be solved. Some of these problems relate to managing patient conditions and deciding what actions to take, whereas others involve problems associated with the nurse's role, collaboration with other providers, and the work environment. The ability to solve patient and setting-related problems is an essential skill to be developed by students. Problem-solving begins with recognizing and defining the problem, gathering data to clarify it further, thinking through possible approaches to use, and evaluating their outcomes. Students faced with patient problems for which they lack understanding and a relevant knowledge base will be impeded in their thinking. This is an important point in both teaching and assessment. When students have an understanding of the problem and possible solutions, they can more easily apply this knowledge to new situations they encounter in the clinical setting.

Past experience with similar problems, either real problems in the clinical setting or hypothetical scenarios in simulation and examples used in teaching, also influences students' skill in problem-solving. Experience with similar problems gives the student a perspective on what to expect in the clinical situation—typical problems the patient may experience and approaches that are usually effective for those problems.

Well-Structured and Ill-Structured Problems

Brookhart and Nitko (2019) defined two types of problems that students may be asked to solve: well structured and ill structured. Well-structured problems provide the information needed for problem-solving; typically, they have one correct solution rather than multiple ones to consider and in general are “clearly laid out” (p. 226). These are problems and solutions that the teacher may have presented in class or online and then asked students about in an assessment. Well-structured problems provide practice in applying concepts learned in class to scenarios and other exemplars but do not require extensive thinking skills.

In contrast, ill-structured problems reflect real-life problems and clinical situations that students encounter in practice. With these situations, the problem may not be clear to the learner, the data may suggest a variety of problems, or there may be an incomplete data set to determine the problem. Along similar lines, the student may identify the problem but be unsure of approaches to take and how to interpret the clinical situation. Some assessment methods may address well-structured problems, assessing understanding of typical problems and approaches. Other methods assess

students' ability to analyze situations to interpret patient needs and concerns, identify possible problems given the data, identify additional data needed, compare and reason through multiple potential approaches, and arrive at an informed judgment as to actions to take (or not) in the situation.

■ Critical Thinking

There has been extensive literature in nursing for many decades about the importance of students developing the ability to think critically. The complexity of patient needs, the extensive amount of information the nurse has to process and analyze in the practice setting, the types of clinical judgments required for providing quality and safe care, and multiple ethical issues faced by the nurse require the ability to think critically. Many definitions of *critical thinking* exist. One way to view critical thinking is that it is the ability to use higher level cognitive skills such as analysis to think through situations and arrive at appropriate judgments and actions; it is being deliberate about thinking (Alfaro-LeFevre, 2017; Papp et al., 2014). To engage in critical thinking, students need a knowledge base to understand the situation, analyze clinical information, make informed decisions, and reflect on the thinking process (Von Colln-Applying & Giuliano, 2017). Students cannot engage in critical thinking if they have not learned about the patient problem or concept earlier.

In the clinical setting, critical thinking enables the student to arrive at sound judgments about patient care. Carrying out assessment; planning care; intervening with patients, families, and communities; and evaluating the effectiveness of interventions—all of these require critical thinking. In the assessment process, important cognitive skills include differentiating relevant from irrelevant data, identifying cues in the data, identifying additional data to collect prior to deciding on the problem, and specifying patient needs and problems based on these data.

Critical thinking also is reflected in the ability to compare possible interventions, considering the evidence, to decide on the best approaches to use in a particular situation (Alfaro-LeFevre, 2017; Facione, 2015). Judgments about the quality and effectiveness of care are influenced by the learner's thinking skills. Nurses, even as they develop their expertise, need to reflect on their clinical reasoning and continue to build their critical thinking skills. Students who demonstrate critical thinking ability:

- Ask questions, are inquisitive, and are willing to search for answers.
- Consider alternate ways of viewing information.
- Offer different perspectives to problems, solutions, and clinical situations.
- Question current practices and express their own ideas about care.

- Extend their thinking beyond the readings, course and clinical activities, and other requirements.
- Are open-minded.

These characteristics are important because they suggest behaviors that are to be developed by students as they progress through the nursing program. They also provide a framework for faculty members to use when assessing whether students have developed their critical thinking abilities.

■ Clinical Judgment

Tanner (2006) viewed clinical judgment as an interpretation of the patient's needs and problems, and decisions on actions to take, or not take, based on the patient's responses. The clinical judgment process includes four aspects: (a) noticing, (b) interpreting, (c) responding, and (d) reflecting. Tanner's model provides a framework for assessing students' thinking in a clinical situation or scenario. Students can be asked to describe what they would expect to find in the situation, what they noticed first, and other data they need (noticing). To assess students' ability to interpret a situation, the teacher can ask them to explain specific data and what they mean, and their priorities of care (interpreting). Another series of questions can explore interventions for the patient, why students would select those interventions or not take any actions, and their rationale (responding). The teacher can ask students to reflect on their experiences with patients and discuss what they would do differently next time (reflecting). In addition to written reflections after clinical experiences, and sharing the thought process used to arrive at judgments, simulations are an effective strategy for helping students develop their clinical judgment skills (Bussard, 2018; Victor, 2017).

■ Context-Dependent Item Sets

In assessing students' cognitive skills, test items and other methods need to meet two criteria. They should (a) introduce *new* information not encountered by students at an earlier point in the instruction and (b) provide data on the thought process used by students to arrive at an answer, rather than the answer alone. Context-dependent item sets may be used for this purpose.

Writing Context-Dependent Item Sets

A basic principle of assessing higher level skills is that the test item or other assessment method has to introduce new or novel material for analysis. Without the introduction of new material as part of the assessment, students may rely on memorization from prior discussion or their readings about how to reason through a clinical situation and decide on actions to take for the situation at hand; they may

simply recall the typical problems and approaches without thinking through other possibilities themselves. In nursing education, this principle is often implemented through clinical scenarios that present a novel situation for students to analyze. Test items that include scenarios in the stem, which students analyze and then answer questions about, are used commonly in nursing education and on licensure and certification examinations. Brookhart and Nitko (2019) referred to these items as *context-dependent item sets* or *interpretive exercises*.

In a context-dependent item set, the teacher presents introductory material that students then analyze and answer questions about. The introductory material may be a description of a clinical situation, patient data, research findings, issues associated with clinical practice, and varied types of scenarios. The introductory material also may include diagrams, photographs, tables, figures, and excerpts from reading materials. Students read, analyze, and interpret the introductory material and then answer questions about it or complete other tasks. One advantage of a context-dependent item set is that it offers an opportunity to present new information for students to analyze that is related to clinical practice. In addition, the introductory material provides the same context for analysis for all students.

The questions asked about the introductory material may be selected- or constructed-response items. With selected-response items such as multiple choice, however, the teacher is not able to assess the underlying thought process used by students in arriving at the answer; their responses reflect instead the outcomes of their thinking. If the goal is to also assess the process students used to think through the situation and decide on an approach, then open-ended items such as short answer and essay would be better items to use.

Interpretive Items on the NCLEX®

On the NCLEX (National Council Licensure Exam), items may include multimedia, such as tables, charts, graphics, and audio, for candidates to interpret and respond to questions. Any of the types of item formats may be used with these, including the standard multiple-choice format and alternate formats (National Council of State Boards of Nursing, 2019). Alternate formats include multiple-response, fill-in-the-blank calculation, ordered-response, hot-spot, and exhibit items. Multiple-response items were presented in Chapter 5, Multiple-Choice and Multiple-Response, and fill-in-the-blank and ordered-response items were discussed in Chapter 6, Short-Answer (Fill-in-the-Blank) and Essay. In a hot-spot item, candidates are asked a question about an image; they answer the question by clicking on the image with the mouse. For example, the candidate might be presented with an image of the chest and asked where to place the stethoscope to listen to heart sounds in the mitral area. In exhibit items, candidates are given a problem, and to answer that problem, they need to read and interpret information in an exhibit. Examples of hot-spot and exhibit items are included later in Exhibit 7.2.

Students should have experience answering these types of questions and other forms of context-dependent items as they progress through a nursing program. Items can be incorporated into quizzes and tests; can be developed for small-group analysis and discussion in class, as out-of-class assignments, and as online activities; and can be analyzed and discussed by students in postclinical conferences. Context-dependent items are also used on nursing certification examinations, and students in graduate programs need to have experience thinking through and answering these types of questions.

Layout

The layout of the context-dependent item set, that is, the way it is arranged on the page, is important so that it is clear to the students which questions relate to the introductory material. Exhibit 7.1 illustrates one way of arranging the material and related items on a page. A heading should be used to indicate the items that pertain to the introductory material, for example, “Questions 1 to 3 refer to the following scenario.” Brookhart and Nitko (2019) suggested that the material for interpretation be centered between the left and right margins of the page so it is readily apparent to the students. If possible, the context and all items pertaining to it should be placed on the same page.

Strategies for Writing Context-Dependent Items

Suggestions follow for writing context-dependent item sets. If the intent is to assess students’ critical thinking or clinical judgment, the introductory material needs to provide sufficient information for analysis without directing the students’ thinking in a particular direction. The first step is to draft the types of questions to be asked about the situation, then to develop a scenario to provide essential information for analysis. If the scenario is designed on the basis of clinical practice, students may be asked to analyze data, interpret the scenario, identify patient problems, decide on nursing interventions, evaluate outcomes of care, and examine ethical issues, among other tasks. Cases, discussed later in this chapter, use a short clinical scenario followed by one or more questions using any type of item.

EXHIBIT 7.1

LAYOUT OF CONTEXT-DEPENDENT ITEM SETS

Questions 1–3 relate to the following scenario:

Scenario (and other types of introductory material) here

1. Item one here
2. Item two here
3. Item three here

The introductory material should be geared to the students' level of understanding and experience. The teacher should check the terminology used, particularly with beginning students. The situation should be of reasonable length without extending the students' reading time unnecessarily.

The questions should focus on the underlying thought process used to arrive at an answer, not on the answer alone. For that reason, short-answer or essay items are frequently used. In some situations, however, the goal may be to assess students' ability to apply concepts or protocols learned in class without any original thinking about them. In this case, scenarios that are clearly laid out for students and questions, such as multiple-choice with one correct answer, are appropriate. Context-dependent items may be incorporated within a test, completed individually or in small groups for formative evaluation, discussed in class or an online environment for instructional purposes, completed during postclinical conferences, or done as out-of-class assignments, either graded or ungraded.

Item sets focusing on assessment of problem-solving ability may ask students to complete the following tasks:

- Identify the problem and alternate problems possible.
- Develop questions for clarifying the problem further.
- Identify assumptions made about the problem and possible approaches or solutions.
- Identify additional data needed for interpreting the situation.
- Differentiate relevant and irrelevant data.
- Propose interventions and summarize supporting evidence.
- Relate knowledge from different sources to the situation to better understand it.
- Evaluate the effectiveness of approaches to solving problems and the outcomes achieved.

The following item set assesses students' skill in identifying and thinking through problems. After reading the introductory situation about the patient, students are asked to identify *all possible* problems and provide data to support them. Other questions ask students about additional data to be collected, again with a rationale for their answer.

Your 8-year-old patient had a closed head injury 4 weeks ago after falling off his bike. You visit him at home and find that he has weakness of his left leg. His mother reports that he is “getting his rest” and “sleeping a lot.” The patient seems irritable during your visit. When you ask him how he is feeling, he tells you, “My head hurts where I hit it.” The mother appears anxious, talking rapidly and changing position frequently.

1. List all possible problems in this situation. For each problem, describe supporting assessment data.
2. What additional data are needed, if any, to decide on these problems? Provide a rationale for collecting this information.
3. What other data would you collect at this time? Why is this information important to your thinking about what to do?

Context-dependent items may focus on actions to be taken in a situation. For this purpose, the teacher can briefly describe a critical event, then ask learners what they would do next. Because the rationale underlying the thinking is as important if not more important than the decision or outcome, students can include an explanation of the thought process they used. For example:

You are caring for a patient with diabetes whose recent urinalysis revealed 4+ ketones and trace leukocytes, and was negative for nitrites and red blood cells.

1. As the nurse practitioner, what would you do next? Why did you choose this action?
2. Which of the laboratory results in the scenario are **most** important in your decision? Provide a rationale for your answer.

On a test with selected-response items, the stem can introduce the critical event or clinical situation, followed by a multiple-choice, multiple-response, or other alternate item format. An example of this type of item using the prior scenario is:

You are caring for a patient with diabetes whose recent urinalysis revealed 4+ ketones and trace leukocytes, and was negative for nitrites and red blood cells. Which of the following actions should the nurse practitioner take next?

- a. Order an ultrasound of the kidneys to rule out subacute renal failure
- b. Check the patient’s blood glucose¹
- c. Order a 24-hour urine test for microalbumin
- d. Check for a history of illicit drug and alcohol use

¹Correct answer.

If the goal is to assess students' ability to think through different decisions possible in a situation, two approaches may be used with the item set. The introductory material (a) may present a situation up to the point of a decision, then ask students to make a decision or (b) may describe a situation and decision and ask whether they agree or disagree with it. For both of these approaches, the students need to provide a rationale for their responses. Examples of these strategies follow:

Your nurse manager on a busy surgery unit asks you to cover for her while she attends a meeting. You find out later that she left the hospital to run an errand instead of attending the meeting.

1. Identify two possible actions you could take in this situation.
2. What would you do? Why?

A patient calls the office to see whether he can receive his flu shot today. He had a cold a few days ago but is feeling better and has returned to work. The nurse instructs the patient to come in for his flu shot.

1. Do you agree or disagree with the nurse's decision?
2. Why or why not?

A 70-year-old patient comes to the urgent care clinic complaining of a bright red-colored spot present in his left eye for 3 days. He has no eye pain, visual changes, or headaches. He is taking one tablet of aspirin a day. Which of the following decisions is most appropriate for this patient?

- a. Report this information immediately to the physician in the clinic.
- b. Get a referral for an ophthalmologist.
- c. Take no action because this condition is benign and will resolve spontaneously.¹
- d. Check to see whether there is an order for an ophthalmic antibiotic.

Often context-dependent item sets are developed based on clinical scenarios. Students can be given material to read and analyze, presented with tables for interpretation, and given images and diagrams with questions to answer. Context-dependent items provide a way for teachers to examine how well students use information and can think through situations. Examples of context-dependent item sets are found in Exhibit 7.2.

EXHIBIT 7.2**SAMPLE CONTEXT-DEPENDENT ITEM SETS AND HOT-SPOT AND EXHIBIT ITEMS****EXAMPLES OF CONTEXT-DEPENDENT ITEM SETS**

Questions 1 and 2 relate to the following situation.

You are unsure about a medication for one of your patients. When you call the pharmacy to learn more about the drug, you discover that the amount ordered is twice the acceptable dose. You contact the attending physician who tells you to “give it because your patient needs that high a dose.”

1. What are your different options at this time? Describe advantages and disadvantages of each.
2. How would you solve this dilemma?

Questions 1–3 relate to the following situation.

Your ventilated patient has his bed elevated 45°. He is being turned, but you notice he is developing a pressure ulcer. Another nurse tells you to lower the head of the bed.

1. Do you agree with that decision? Why or why not?
2. What would you do?
3. What evidence supports your reasoning?

Questions 1–4 relate to the following scenario.

A 1-month-old girl is brought to the pediatrician’s office for a well-baby checkup. You notice that she has not gained much weight over the last month. Her mother explains that the baby is “colicky” and “spits up a lot of her feeding.” There is no evidence of projectile vomiting and other gastrointestinal symptoms. The baby has a macular-type rash on her stomach, her temperature is normal, and she appears well hydrated.

1. Describe possible problems this baby may be experiencing. What additional data would you collect next? Why?
2. What would you do in this situation? What evidence supports those interventions?
3. Specify outcomes for evaluating the effectiveness of the interventions you selected.
4. What information presented in this situation is not relevant to your decision-making? Why?

A patient is seen in the clinic for a severe headache, nausea, and vomiting, which has lasted for 3 days. His blood pressure (BP) is 230/115 mmHg and heart rate is 94 beats per minute (BPM). He takes atenolol (Tenormin) at home. The patient is ordered an intravenous (IV) sodium nitroprusside infusion. At 10 minutes after the infusion, the nurse checks the patient’s BP, and it is the same. His mean arterial pressure (MAP) is 150 mmHg. The nurse should:

- a. report this to the physician or other advanced practice provider.
- b. stop the IV infusion and monitor BP.
- c. increase the dose of the sodium nitroprusside per order.*
- d. maintain the infusion and recheck BP and MAP in 30 minutes.

(continued)

EXHIBIT 7.2**SAMPLE CONTEXT-DEPENDENT ITEM SETS AND HOT-SPOT AND EXHIBIT ITEMS
(continued)**

Use this table to answer the questions.

	MEN		WOMEN		
Importance Ratings	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>t</i>
Able to call RN with questions	4.23	.93	4.92	.95	2.76 ^a
Have RN teach illness, medications, treatment options	4.47	.79	4.40	.90	.57
Have RN teach health promotion	4.35	.90	4.00	1.1	2.51 ^a

Note: ^a $p < .01$. *M*, mean. *SD*, standard deviation.

Based on the data presented in the table, which of the following conclusions is accurate?

1. Health-promoting activities were more important to men than to women.¹
2. It was more important to men to be able to call an RN with questions after a visit.
3. Men valued teaching by the RN more than women.
4. Teaching about health was more important to women than men.

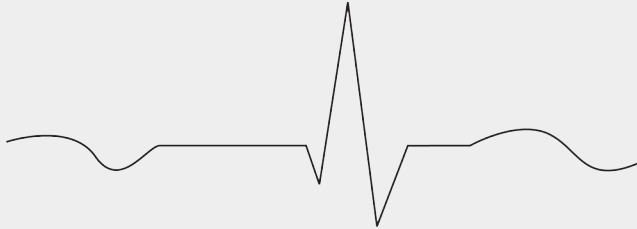
An 82-year-old woman is scheduled for an open reduction and internal fixation of her right femur. She rates her pain at 4 on a 1–10 scale and has not received any preoperative sedation. Three days ago on admission, an indwelling urinary catheter was inserted and it is now draining clear, light yellow urine. While transferring the patient to the operating room (OR) bed, the perioperative RN notices a reddened area on her sacrum, approximately the size and shape of a silver dollar; the skin is intact with no blistering apparent, and it feels warm to touch. The patient is positioned supine on the fracture OR bed with the left leg supported in a padded stirrup. Her right arm is secured at her side, and the left arm is extended at a 45° angle on a padded arm board.

- | | | |
|--------|----|--|
| T F | 1. | According to the National Pressure Ulcer Advisory Panel definitions of pressure ulcers, the perioperative RN should document the lesion on the patient's sacrum as stage I. |
| T F | 2. | This patient is at an increased risk for inability to give informed consent for the surgical procedure. |
| T F | 3. | The patient's position on the OR bed increases her risk of compression and stretch injury to the left sciatic nerve. |
| T F | 4. | According to the Centers for Disease Control and Prevention guideline for prevention of catheter-associated urinary tract infections, this patient has an increased risk of such an infection. |

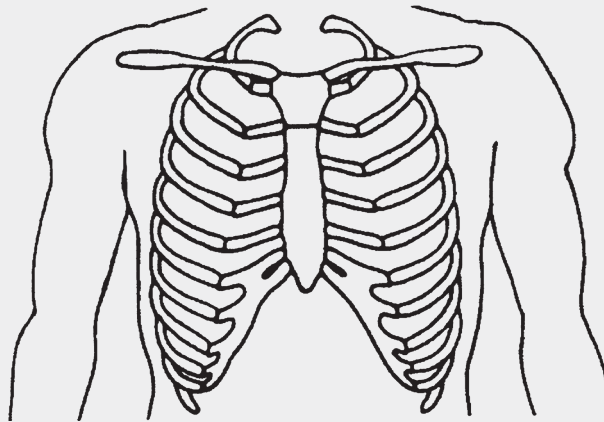
(continued)

EXHIBIT 7.2**SAMPLE CONTEXT-DEPENDENT ITEM SETS AND HOT-SPOT AND EXHIBIT ITEMS**
(continued)**EXAMPLES OF HOT-SPOT ITEMS**

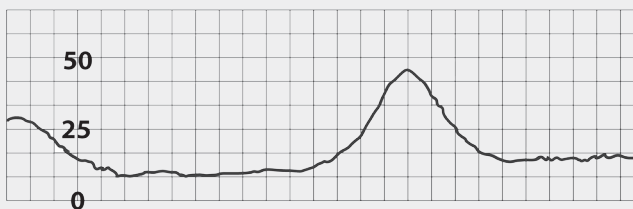
On the following EKG, mark the area of the ST segment.



Your patient has an aortic stenosis. Mark the spot where you would place the stethoscope to best hear the murmur.



Click on the area that represents the beginning of the contraction.



(continued)

EXHIBIT 7.2

SAMPLE CONTEXT-DEPENDENT ITEM SETS AND HOT-SPOT AND EXHIBIT ITEMS
(continued)

EXAMPLES OF AN EXHIBIT ITEM

You are caring for a 2-day postpartum patient with a history of lupus. She had an emergency cesarean delivery at 36 weeks' gestation and during the delivery began to bleed, leading to hypovolemic shock. She received blood and fluid replacements. In morning report, you are told that the patient is stable but drowsy. The following information is in the medical record:

VITAL SIGNS					
DATE	TIME	TEMP	PULSE	RESP	B/P
2/1	0600				
	1000	99.4	80	28	102/52
	1400				
	1800		88		
	2200				
2/2	0200				
	0600	99.8	88	30	124/60
	1000				
	1400		100	30	
	1800				
	2200	100.2	120	32	98/56
	0200				

Which of the following information is most important to collect next?

- Appearance of the incision site
- Breath sounds
- Type of vaginal discharge
- Urinary output for the past 24 hours¹

A patient is admitted with auditory and visual hallucinations and a history of depression. Her hallucinations have increased in frequency and severity in the last 2 weeks. The patient is on paroxetine (Paxil) for depression, and she reports using herbal supplements for her hallucinations. The nurse reviews these data in the electronic medical record:

Vital signs: Temperature: 98.2°F, respiratory rate: 24 breaths per minute, heart rate: 90 beats per minute, blood pressure: 132/86 mmHg

Physical examination: Head symmetrical, dull affective responses, pupils equal and reactive to light and accommodation, unable to fully extend arms, has ataxic gait

The nurse should suspect that these assessment findings are related to:

- Parkinson's disease.
- extrapyramidal side effects.¹
- worsening depression.
- interaction of herbals and Paxil.

■ Assessment Methods for Higher Level Cognitive Skills

Although context-dependent item sets provide one means of assessing higher level cognitive skills, other methods are available for this purpose. Those alternate approaches include cases, case study, and unfolding cases; discussion; debate; video clips; short written assignments; and varied clinical evaluation methods, which are presented in Chapter 14, Clinical Evaluation Methods. Many of the assessment methods described in this section of the chapter also may be used for clinical evaluation.

Cases, Case Study, and Unfolding Cases

With cases, students analyze a clinical scenario and answer related questions. The focus might be on interpreting patient needs and problems, identifying additional data to collect, applying concepts and readings to the case, examining the case from different points of view, and identifying interventions. Cases engage students in thinking through problems they might encounter in their clinical practice and are effective for developing higher level cognitive skills (Cleveland, Carmona, Paper, Solis, & Taylor, 2015; Hong & Yu, 2019; Li, Ye, & Chen, 2019; Thistlethwaite et al., 2012; Vacek & Liesveld, 2019). When using cases, the scenarios can be short, providing only essential information about the clinical situation, in contrast to case studies, which are longer and offer more detail.

Cases work well for group analysis and discussion, either in class or online as small-group activities or in postclinical conference. Cleveland et al. (2015) developed web-delivered cases using e-learning authoring software and delivered through their learning management system. With electronic case scenarios, technology can be added to improve the reality of the clinical situation and opportunities for using higher level thinking skills. Shellenbarger and Robb (2015) described embedding hyperlinks and adding podcasts, graphics, images, and video clips to case scenarios. Allowing the clinical situation to unfold enables students to think through the situation, identify priorities, and cluster information, all part of clinical reasoning. In groups, students can critique each other's thinking; compare different interpretations of the problem, interventions, and decisions possible; and learn how to arrive at a group consensus. Used as a small-group activity, the case method is more easily evaluated for formative than summative purposes. Exhibit 7.3 presents examples of a short case, case study, and unfolding case. These could be used for instruction or for assessment, particularly formative, allowing students to critique each other's ideas and receive feedback on their own. Students' analyses of cases and their responses to questions also can be evaluated and scored similarly to an essay item.

EXHIBIT 7.3**SAMPLE CASE, CASE STUDY, AND UNFOLDING CASE****CASE**

A 92-year-old man is brought to the emergency department by his son. The patient seems to be dragging his right leg and has slurred speech. His blood pressure is 220/110.

1. What are possible problems this patient might be experiencing?
2. What additional data will you collect from the son, and why is this information important to confirming the problem?

CASE STUDY

A 20-year-old woman has had abdominal pain for the past 2 weeks. Some mornings she has vomiting, but today she complains mainly of severe abdominal cramps and nausea. She has lost 8 lb since last week and has no appetite. She reports having diarrhea for the past few days. She has no masses that you can feel although she complains of increased pain with even a slight touching of her abdominal area. Her vital signs are normal.

Her mother, who brought her to the office today, reports that the patient has always been healthy and has had no prior illnesses except for colds and an occasional flu. She lives with both parents and her younger brother, and she is a student at the local college.

1. What are possible problems that this patient might have? What data would you collect to narrow down your list of problems?
2. What laboratory tests would you expect to be ordered? Why?
3. As you talk with the patient's mother, you learn that the family was on a cruise a few weeks ago, but no one "got sick on the cruise." How might this new information influence your thinking about the patient's possible problems?
4. Considering only the data presented in the case, develop a care plan to meet the patient's current needs. Provide a rationale for each intervention in your plan with a summary of the evidence.

UNFOLDING CASE

You are making a home visit to see a 71-year-old woman who has a leg ulcer that began after she fell. The patient is coughing and wheezing; she tells you she "feels terrible."

1. What additional data would you collect in the initial assessment? Why?
2. What actions would you take during this home visit? Provide a rationale.

In 3 days, you visit this patient again. She has increased shortness of breath, more fatigue, and a pale color, and she seems cyanotic around her mouth.

1. Does this new information change your impression of her problems? Why or why not?
2. List priority problems for this patient with a brief rationale.
3. What will you report to the physician when you call?

The patient recovers from that episode, and you are able to visit her one more time. At this last visit, she is still short of breath but otherwise seems improved. Write your discharge summary for this patient.

A case study provides a hypothetical or real-life situation for students to analyze and then arrive at varied decisions. Case studies are more comprehensive than the introductory material presented with the case method (Exhibit 7.3). With case studies, students are able to provide detailed and in-depth analyses and describe the evidence on which their conclusions are based. The case study also provides a means for students to apply relevant concepts from class and from their readings. A case study may be completed as an individual assignment and assessed similarly to other written assignments as long as the students provide a rationale for their decisions. The results of the case analysis may be presented orally for group critique and feedback.

One other method to use to assess higher level learning is unfolding cases, which provide a means of simulating a patient situation that changes over time. Rather than writing one short case scenario or a more comprehensive one with background information as in a case study, unfolding cases describe changes in a patient's condition, clinical situation, or setting of care similar to what might occur with an actual patient (see Exhibit 7.3). An unfolding case can be "intentionally unpredictable" to reflect the real world of clinical practice and to foster development of students' clinical reasoning skills (Smallheer, 2016, p. 7). Unfolding case studies are valuable for guiding students in applying concepts and knowledge to clinical practice and developing their thinking skills. This method also enables students to identify the most salient aspects of a case, develop their clinical judgment, and begin to think like a nurse (Benner, Sutphen, Leonard, & Day, 2010; Bowman, 2017; Zook, Hulton, Dudding, Stewart, & Graham, 2018). Smallheer (2016) described using a reverse unfolding case, in contrast to an unfolding case in which a scenario is presented to students. In a reverse case study, students create the scenario. This type of unfolding case would work well for formative assessment.

Discussion

Discussions with students individually and in small groups are an important strategy for assessing students' understanding and ability to apply their knowledge to clinical practice, critical thinking, and clinical judgment. In a discussion, the teacher has an opportunity to ask questions about students' thinking and the process they used for arriving at decisions and positions on issues. Discussions may be impromptu, used for formative evaluation, or structured by the teacher to provide a context and questions to which students respond. Use of discussion for assessing cognitive skills, however, requires careful questioning with a focus on the thinking used by students to arrive at answers. In these discussions, the teacher can ask students about possible decisions, implications of options they considered, and different points of view in the situation. Discussions of clinical situations, real or hypothetical, allow the teacher to think aloud about data to collect and why this information is important, potential problems to consider, and different approaches that might be used. Students can

share how they would think through a situation prior to acting and why, and receive feedback from the teacher.

The difficulty level of questions asked is significant; the teacher should avoid a predominance of factual questions and focus instead on higher level questions. With factual questions, students recall facts and specific information about the problem, clinical situation, or issue being discussed. For example, factual questions are: “What is dyspnea?” and “What are neutrophils?” Clarifying questions explore students’ understanding of the topic. Examples of clarifying questions are: “Tell me about the relationship between your patient’s shortness of breath and her cardiac problems” and “Why is it important to check the patient’s white blood cell count?” For those questions, students explain their answers using their own words. Higher level questions cannot be answered by memory alone and require an evaluation or a judgment of the situation. Examples of higher level questions are: “What are similarities and differences between the assessment and diagnoses for Mrs. S and the patient you had last week?” and “Which pain interventions would you propose for this patient? Why did you decide on these interventions rather than the others?”

Questions for discussions should be sequenced from a low to a high level, beginning with factual questions to assess students’ knowledge of relevant facts and concepts and their ability to apply them to the situation, problem, and issue, and progressing to questions that assess students’ thinking and clinical judgments. The taxonomy can be used as a framework for developing questions for discussions focusing on higher level thinking. With this schema, low-level questions require students to remember facts and explain their answers. With these questions, the teacher can assess students’ knowledge and whether they understand the information. These questions provide a good opportunity for the teacher to give feedback and fill gaps in learning. Higher level questions would focus on applying, analyzing, evaluating, and creating. This taxonomy of the cognitive domain was described and examples of each level were provided in Chapter 1, Assessment and the Educational Process.

If discussions are to be geared toward assessment of higher level thinking, nurse educators need to be aware of the level of questions they ask students. Although it is important to assess students’ understanding of a clinical situation, the teacher’s questions should move beyond that. Students should be asked to describe what they would expect to find in a clinical situation, what they noticed first when they observed the patient, and what other data they need to collect. With carefully planned questions from the educator, students can be asked to explain specific data and what they mean, their priorities of care and why, interventions and evidence for use, and how they would care for the patient differently next time. With those types of questions, teachers can more easily assess students’ thinking and clinical judgment in a clinical situation. Open-ended questions with multiple possible answers are a good type of question for formative assessment of students’ thinking (Brookhart, 2014).

With a logical sequence of questions, students can analyze complex issues, examine alternate points of view, and draw generalizations across different content areas. However, these outcomes will not be achieved without carefully thought-out questions by the teacher.

Debate

Debate provides an effective mechanism for assessing students' ability to analyze issues in depth, consider alternative points of view, and formulate and present a position. The process of analyzing the issue for the debate, considering alternative viewpoints, developing a sound position, and preparing persuasive arguments promotes critical thinking skills. Hartin, Birks, Bodak, Woods, and Hitchins (2017) suggested that debates can be an effective way to foster the type of creative and forward-thinking nurses needed in healthcare

The focus in evaluating a debate should be on the strength of the argument developed and presented to the group. Areas to consider in evaluating debates include:

1. Clarity and comprehensiveness of the analysis of the issue
2. Rationale developed for the position taken, including use of the literature and available research
3. Consideration of alternative positions
4. Clarity of responses to the opposing side
5. Organization and development of the argument
6. Degree of persuasiveness in presenting the argument
7. Presentation skills, including keeping the audience interested and focused, presenting the information logically and clearly, and keeping within the allotted time frame

Depending on the size of the class, not all students may be able to participate in the debate, but they can all learn from it. Presenting the debate allows students to practice their oral communication skills, make an argument about a position, and speak to a group (Bradshaw, 2017).

Video Clips

Video clips, YouTube, virtual reality, and other types of multimedia may be used to present a scenario for assessing higher level learning. Multimedia adds to the reality of the situation. Any type of technology may be used for this purpose. For example, images, video and audio clips, virtual reality, and many other educational and computer technologies can be used to develop real-life scenarios for students to analyze, interact with, and discuss. There is a wealth of resources on the web for

presenting scenarios and other situations for teaching and assessing higher level cognitive skills. These can be integrated easily within an online learning environment, and students can work individually or in groups to analyze them.

Short Written Assignments

Evaluation of written assignments is presented in Chapter 9, Assessment of Written Assignments. For the purposes of assessing higher level thinking and other cognitive skills, however, these assignments should reflect additional principles. Assignments for this purpose should be short and require students to think critically about the topic. With term papers and other long assignments, students often summarize the literature and report on the ideas of others, rather than thinking about the topic themselves. Short written assignments, in contrast, provide an opportunity for students to express their thinking in writing and for teachers to give prompt and specific feedback to them on their reasoning.

Students should have clear directions as to what to write about and the expected length of the assignment. Assignments can be planned throughout a course and level in a nursing program so that they build on one another, helping students to develop gradually their thinking and writing skills (Oermann et al., 2015). Beginning assignments can ask students to describe a problem, a clinical situation, or an issue and how they would think through the situation and respond to it. In later assignments, students can develop arguments to support their own positions about issues.

Examples of written assignments for assessing higher level cognitive skills, appropriate for either formative or summative evaluation, include short papers that:

- Analyze different data sets.
- Compare problems and alternative interventions that could be used.
- Analyze issues.
- Analyze different points of view, perspectives, and positions on an issue.
- Compare a student's own and others' positions on an issue or topic.
- Present evidence on which their reasoning is based.
- Analyze clinical judgments and possible alternatives given the same evidence.
- Present an argument to support a position.

■ Summary

This chapter provided a framework for assessing higher level learning skills among nursing students. The ability to solve patient and other types of problems is an essential ability to be developed by nursing students. The nurse continually makes

decisions about problems, interventions, possible alternatives, and the best approach to use in a particular clinical situation.

In assessing higher level cognitive skills, as a basic principle the teacher introduces new or novel material for analysis. Without the introduction of new material as part of the assessment, students may rely on memorization of content from prior discussion or their readings about that patient problem or clinical situation; they may simply recall the typical way of addressing the problem without thinking through alternative possibilities themselves. As a result, an essential component of this assessment is the introduction of new information not encountered by the student at an earlier point in the instruction. In nursing, this is frequently accomplished by developing scenarios that present a novel situation for student analysis. These items are referred to as *context-dependent item sets* or *interpretive exercises*.

In a context-dependent item set, the teacher presents introductory material that students then analyze and answer questions about. The introductory material may be a description of a clinical situation, patient data, research findings, issues associated with clinical practice, and tables, among other types. Students read, analyze, and interpret this material and then answer questions about it or complete other tasks.

Other methods for assessing cognitive skills in nursing were presented in the chapter: cases and case study, unfolding cases, discussions, debate, video clips and other types of multimedia, and short written assignments. In addition to these strategies, clinical evaluation methods that provide for an assessment of cognitive skills are presented in Chapter 14, Clinical Evaluation Methods.

■ References

- Alfaro-LeFevre, R. (2017). *Critical thinking, clinical reasoning, and clinical judgment: A practical approach* (6th ed.). Philadelphia, PA: Elsevier.
- Benner, P. E., Sutphen, M., Leonard, V., & Day, L. (2010). *Educating nurses: A call for radical transformation*. San Francisco, CA: Jossey-Bass.
- Bowman, K. (2017). Use of online unfolding case studies to foster critical thinking. *Journal of Nursing Education*, 56, 701–702. doi:10.3928/01484834-20171020-13
- Bradshaw, M. J. (2017). Debate as a teaching strategy. In M. J. Bradshaw & B. L. Hultquist (Eds.), *Innovative teaching strategies in nursing and related health professions* (7th ed., pp. 189–197). Burlington, MA: Jones & Bartlett Learning.
- Brookhart, S. M. (2014). *How to design questions and tasks to assess student thinking*. Alexandria, VA: ASCD.
- Brookhart, S. M., & Nitko, A. J. (2019). *Educational assessment of students* (8th ed.). New York, NY: Pearson Education.
- Bussard, M. E. (2018). Evaluation of clinical judgment in prelicensure nursing students. *Nurse Educator*, 43, 106–108. doi:10.1097/nne.0000000000000432
- Cleveland, L. M., Carmona, E. V., Paper, B., Solis, L., & Taylor, B. (2015). Baby Boy Jones interactive case-based learning activity: A web-delivered teaching strategy. *Nurse Educator*, 40, 179–182. doi:10.1097/nne.0000000000000129

- Facione, P. A. (2015). *Critical thinking: What it is and why it counts*. Hermosa Beach, CA: Measured Reasons.
- Hartin, P., Birks, M., Bodak, M., Woods, C., & Hitchins, M. (2017). A debate about the merits of debate in nurse education. *Nurse Education in Practice*, 26, 118–120. doi:10.1016/j.nepr.2017.08.005
- Hong, S., & Yu, P. (2017). Comparison of the effectiveness of two styles of case-based learning implemented in lectures for developing nursing students' critical thinking ability: A randomized controlled trial. *International Journal of Nursing Studies*, 68, 16–24. doi:10.1016/j.ijnurstu.2016.12.008
- Li, S., Ye, X., & Chen, W. (2019). Practice and effectiveness of “nursing case-based learning” course on nursing student's critical thinking ability: A comparative study. *Nurse Education in Practice*, 36, 91–96. doi:10.1016/j.nepr.2019.03.007
- National Council of State Boards of Nursing. (2019). *2019 NCLEX® test plan*. Chicago, IL: Author.
- Oermann, M. H., Leonardelli, A. K., Turner, K. M., Hawks, S. J., Derouin, A. L., & Hueckel, R. M. (2015). Systematic review of educational programs and strategies for developing students' and nurses' writing skills. *Journal of Nursing Education*, 54, 28–34. doi:10.3928/01484834-20141224-01
- Papp, K. K., Huang, G. C., Lauzon Clabo, L. M., Delva, D., Fischer, M., Konopasek, L., ... Gusic, M. (2014). Milestones of critical thinking: A developmental model for medicine and nursing. *Academic Medicine*, 89, 715–720. doi:10.1097/acm.0000000000000220
- Shellenbarger, T., & Robb, M. (2015). Technology-based strategies for promoting clinical reasoning skills in nursing education. *Nurse Educator*, 40, 79–82. doi:10.1097/NNE.0000000000000111
- Smallheer, B. A. (2016). Reverse case study: A new perspective on an existing teaching strategy. *Nurse Educator*, 41, 7–8. doi:10.1097/NNE.0000000000000186
- Tanner, C. A. (2006). Thinking like a nurse: A research-based model of clinical judgment. *Journal of Nursing Education*, 45, 204–211.
- Thistlethwaite, J. E., Davies, D., Ekeocha, S., Kidd, J. M., MacDougall, C., Matthews, P., ... Clay, D. (2012). The effectiveness of case-based learning in health professional education. A BEME systematic review: BEME Guide 23. *Medical Teacher*, 34, e421–e444. doi:10.3109/0142159X.2012.680939
- Vacek, J., & Liesveld, J. (2019). Teaching concepts to nursing students using model case studies, the Venn diagram, and questioning strategies. *Nursing Education Perspectives*. Advance online publication. doi:10.1097/01.Nep.0000000000000514
- Victor, J. (2017). Improving clinical nursing judgment in prelicensure students. *Journal of Nursing Education*, 56, 733–736. doi:10.3928/01484834-20171120-05
- Von Colln-Appling, C., & Giuliano, D. (2017). A concept analysis of critical thinking: A guide for nurse educators. *Nurse Education Today*, 49, 106–109. doi:10.1016/j.nedt.2016.11.007
- Zook, S. S., Hulton, L. J., Dudding, C. C., Stewart, A. L., & Graham, A. C. (2018). Scaffolding interprofessional education: Unfolding case studies, virtual world simulations, and patient-centered care. *Nurse Educator*, 43, 87–91. doi:10.1097/nne.0000000000000430

TEST CONSTRUCTION AND PREPARATION OF STUDENTS FOR NCLEX® AND CERTIFICATION EXAMINATIONS

One of the outcomes of prelicensure nursing programs is for graduates to pass an examination that measures their knowledge and competencies to engage in safe and effective nursing practice. At the entry level for professional nursing, graduates take the NCLEX-RN® (National Council Licensure Examination for Registered Nurses) or, if graduating from a practical or vocational nursing program, they take the NCLEX-PN® (National Council Licensure Examination for Practical Nurses). Certification validates knowledge and competencies for professional practice in a specialized area of nursing. As part of this process, nurses may take certification examinations that assess their knowledge and skills in a nursing specialty such as critical care or pediatric nursing. Other certification examinations measure knowledge and competencies for advanced practice, for teaching, and for administrative roles. As students progress through a nursing program, they should have experience with tests that are similar to and prepare them for taking licensure and certification examinations when they graduate.

Because the focus of the NCLEX and most certification examinations is on nursing practice, the other advantage to incorporating items of these types in teacher-made tests is that it provides a way of assessing whether students can apply their theoretical learning to clinical situations. Teachers can develop items that present new and complex clinical situations for students to critically analyze. Items can focus on collecting and analyzing data, setting priorities, selecting interventions, and evaluating outcomes; providing patient-centered care; communicating with patients, their families, and other healthcare providers; documenting care; teaching; and other processes fundamental to nursing practice. This type of testing is a means of assessing higher and more complex levels of learning and provides essential practice before students encounter similar questions on licensure and certification examinations.

The chapter includes examples of items written at different cognitive levels, thereby avoiding tests that focus only on recall and memorization of facts, and sample stems

that can be used for clinically oriented test items. The types of items presented in this chapter are similar to those found on the NCLEX and many certification tests. By incorporating items of these types on tests in nursing courses, teachers help students acquire experience with this type of testing as they progress through the program, preparing them for taking licensure and certification examinations as graduates. The reader should keep in mind that Chapter 7, *Assessment of Higher Level Learning*, presented other ways of assessing higher level learning.

The chapter begins with an explanation of the NCLEX test plans in use at the time this edition of the book was prepared (2019). NCLEX and certification examinations are updated every few years to keep current with practice. Based on the most recent practice analysis, which documented the complex decisions nurses make when caring for patients, the National Council of State Boards of Nursing (NCSBN) is exploring use of innovative types of test items to evaluate clinical judgment—the Next Generation NCLEX project (NCSBN, 2019a). Readers are advised to refer to the NCLEX and certification test plans in use when developing examinations and quizzes for their own nursing students.

■ NCLEX Test Plans

In the United States and its territories, graduates of nursing programs cannot practice as RNs or as practical nurses (PNs) or vocational nurses (VNs) until they have passed a licensure examination. These examinations are developed by the NCSBN based on extensive analyses of the practice requirements of RNs and licensed practical nurses (LPNs) or vocational nurses (LVNs). Items are piloted and tested extensively to ensure they are valid, reliable, and legally defensible, including an analysis of potential biases such as those related to ethnicity and gender (Woo & Dragan, 2012). The licensure examination results then are used by the state boards of nursing as one of the requirements for practice in that state or territory.

■ NCLEX-RN Test Plan

In developing the NCLEX-RN, the NCSBN conducts an analysis of the current practice of newly licensed RNs across clinical areas and settings. This is a continuous process allowing the licensure examination to stay current with the knowledge and competencies needed by entry-level nurses. To ensure that the NCLEX-RN measures the essential competencies for safe and effective practice by a newly licensed RN, the NCSBN reviews the test plan or blueprint every 3 years (NCSBN, 2018). For the most recent revision of the test plan, 2,275 newly licensed RNs prioritized how frequently they performed 142 nursing activities and rated the overall importance of each activity. Findings are used to develop the NCLEX-RN test plan and items on the test.

Client Needs

Test items on the NCLEX-RN are categorized by client needs: (a) safe and effective care environment, (b) health promotion and maintenance, (c) psychosocial integrity, and (d) physiological integrity. Two of the categories, safe and effective care environment and physiological integrity, also have subcategories. The client needs represent the content tested on the examination.

Safe and Effective Care Environment

In the safe and effective care environment category, two subcategories of content are tested on the NCLEX-RN: (a) management of care and (b) safety and infection control. In the management of care subcategory, the test items focus on providing and directing nursing care that enhances care delivery to protect clients and healthcare providers. Examples of content tested in this category include advance directives; advocacy; assignment, delegation, and supervision; case management; collaboration with the interdisciplinary team; concepts of management; confidentiality/information security; continuity of care; establishing priorities; ethical practice; informed consent; legal rights and responsibilities; and performance improvement (quality improvement) among others (NCSBN, 2019b). In the NCLEX-RN, 17% to 23% of the items assess management of care.

In the safety and infection control subcategory, test items focus on prevention of accidents, emergency response planning, ergonomic principles, handling hazardous and infectious materials, reporting incidents and irregular occurrences, safe use of equipment, standard precautions, and use of restraints, among others (NCSBN, 2019b). Between 9% and 15% of the items on the NCLEX-RN relate to safety and infection control.

Health Promotion and Maintenance

The second category of client needs is health promotion and maintenance. Between 6% and 12% of the items on the NCLEX-RN relate to health promotion and maintenance. There are no subcategories of needs. Examples of content tested in this category are the aging process, ante/intra/postpartum and newborn care, developmental stages and transitions, health promotion and disease prevention, lifestyle choices, physical assessment techniques, and others (NCSBN, 2019b).

Psychosocial Integrity

The third category of client needs, psychosocial integrity, also has no subcategories. This category focuses on nursing care that promotes the emotional, mental, and social well-being of clients experiencing stressful events, and the care of patients with acute and chronic mental illness. Examples of content tested include abuse, behavioral

interventions, coping, crisis intervention, cultural awareness and influences on health, end-of-life care, grief and loss, mental health, sensory and perceptual alterations, and therapeutic communication and environment (NCSBN, 2019b). Six percent to 12% of the items on the NCLEX-RN ask questions about psychosocial integrity.

Physiological Integrity

The final client needs category, physiological integrity, is a significant content area tested on the NCLEX-RN. Items in this category focus on nursing care that promotes physical health and comfort, reduces risk potential, and manages health alterations. Four subcategories of content are examined by these items on the NCLEX-RN:

1. *Basic care and comfort*: In this area, items focus on comfort measures and assistance with activities of daily living. Related content includes assistive devices, elimination, mobility and immobility, nonpharmacological comfort interventions, nutrition and oral hydration, personal hygiene, and rest and sleep. Six percent to 12% of the items are on basic care and comfort.
2. *Pharmacological and parenteral therapies*: Items focus on adverse effects, contraindications, side effects, and interactions; blood and blood products; calculating dosages; central venous access devices; medication administration; parenteral/intravenous therapy; pharmacological pain management; and total parenteral nutrition. More test items are included on the NCLEX-RN in this subcategory than the others in the physiological integrity category. Between 12% and 18% of the items relate to pharmacological and parenteral therapies.
3. *Reduction of risk potential*: The content in this subcategory relates to measures for reducing the risk of developing complications or health problems. For example, items relate to diagnostic tests; laboratory values; potential for complications from diagnostic tests, treatments, and procedures, and from surgical procedures; and system-specific assessments, among others. In the test plan, 9% to 15% of the items relate to these content areas.
4. *Physiological adaptation*: The last subcategory includes nursing care of patients with acute, chronic, or life-threatening physical health problems. Between 11% and 17% of the test items relate to physiological adaptation. Sample content areas are alterations in body systems, fluid and electrolyte imbalances, hemodynamics, management of illness and medical emergencies, pathophysiology, and unexpected responses to therapies (NCSBN, 2019b).

Integrated Processes

Five processes that are fundamental to nursing practice are integrated throughout each of the categories of the test plan: (a) nursing process, (b) caring, (c) communication

and documentation, (d) teaching and learning, and (e) culture and spirituality (NCSBN, 2019b). Thus, there can be test items on teaching patients and the nurse's ethical and legal responsibilities in patient education as part of the management of care subcategory, teaching nursing assistants about the use of restraints in the safety and infection control subcategory, health education for different age groups in the health promotion and maintenance category, and teaching about diagnostic tests in the reduction of risk potential subcategory. The other processes are integrated similarly throughout the test plan. Many of the items on the NCLEX examinations are developed based on clinical scenarios. Those scenarios can involve any age group of patients in acute care hospitals, long-term care, community health, or other types of settings.

Cognitive Levels

The taxonomy for the cognitive domain is used for developing and coding items on the NCLEX-RN (NCSBN, 2019b). This taxonomy was presented in Chapter 1, Assessment and the Educational Process. The majority of items are at the application (applying) and higher cognitive levels (NCSBN, 2019b). This has implications for testing in prelicensure nursing education programs. Faculty members should avoid preparing only items that require remembering and understanding facts and specific information on their tests. Although some low-level questions are essential to assess students' knowledge, test items also need to ask students to *use* their knowledge and analyze data and clinical situations to decide on approaches to take. Test blueprints can be developed to list not only the content and number of items in each content area but also the level of cognitive complexity at which items should be written. An example of a blueprint of this type was provided in Exhibit 3.3 in Chapter 3, Planning for Testing.

■ NCLEX-PN Test Plan

The test plan for the NCLEX-PN is developed and organized similarly to the RN examination. In the most recent practice analysis for this test plan, LPN/LVNs who were newly licensed were asked how frequently they performed 151 nursing activities and the importance of those activities (NCSBN, 2019c). The activities were then used as the framework for the development of the test plan for the PN examination.

The test plan is structured around client needs and integrated processes fundamental to the practice of practical and vocational nursing. The same four client needs categories are used for the NCLEX-PN with differences in some of the subcategories, related content, and percentage of items in each category and subcategory. In the safe and effective care environment category on the NCLEX-PN, 18% to 24% of the items are on coordinated care, and 10% to 16% are on safety and infection control. Between 6% and 12% of the items assess the second category of client needs: health promotion and maintenance. Nine percent to 15% of the items

on the NCLEX-PN address psychosocial integrity. The last category of client needs, physiological integrity, has four subcategories similar to the RN test plan: basic care and comfort (7%–13% of the items), pharmacological therapies (10%–16% of the items), reduction of risk potential (9%–15% of the items), and physiological adaptation (7%–13% of the items; NCSBN, 2016). Five processes are integrated throughout the test: (a) the clinical problem-solving process (nursing process), (b) caring, (c) communication and documentation, (d) teaching and learning, and (e) culture and spirituality. Items are developed at all cognitive levels with the majority written at the application or higher levels of cognitive abilities, similar to the NCLEX-RN test plan (NCSBN, 2016).

■ Types of Items on the NCLEX Examinations

The NCLEX examinations contain multiple-choice items and alternate item formats. Earlier chapters described how to construct each type of item used on the NCLEX: multiple-choice (Chapter 5, Multiple-Choice and Multiple-Response); the alternate formats of multiple-response (Chapter 5, Multiple-Choice and Multiple-Response), fill-in-the-blank (for calculations), and ordered response (Chapter 6, Short-Answer (Fill-in-the-Blank) and Essay); and hot-spot and chart or exhibit (Chapter 7, Assessment of Higher Level Learning). Some items include audio and the candidate listens to an audio clip and selects a response; other items have graphics. Any of the types of items might include a table, a chart, an image, or sound as part of the item. These items with multimedia have the capacity to assess higher levels of thinking and do so more authentically than a text-based item.

The NCLEX-RN Detailed Test Plan (National Council of State Boards of Nursing, 2016, 2019b) provides valuable information about the practice activities used for developing the items and content areas assessed in each of the categories and subcategories on the examination. As described earlier, the NCSBN analyzes the current practices of newly licensed RNs and PNs or VNs across clinical specialties and settings and the knowledge needed for safe and effective practice. This analysis identifies nursing activities that are used frequently by entry-level nurses and are important to ensure patient safety. Development of the NCLEX using these practice activities provides evidence of validity as a measure of entry-level nursing practice.

The NCLEX-RN Detailed Test Plan includes a list of the activity statements and related content for each category and subcategory. This information is of value in developing items for tests in a nursing program. For example, in the safety and infection control subcategory, the activity statements describe the practices that RNs use to protect clients and healthcare personnel from health and environmental hazards. An example of one of these activity statements is, “Apply principles of infection control (e.g., hand hygiene, aseptic technique, isolation, sterile technique, universal/standard precautions)” (NCSBN, 2019b, p. 15). A sample test item is also provided with each category and subcategory.

■ Administration of NCLEX Examinations

The NCLEX is administered to candidates by computerized adaptive testing (CAT). The CAT model is such that each candidate's test is assembled interactively as the person is answering the questions. Each item on the NCLEX has a predetermined difficulty level. As each item is answered, the computer reestimates the candidate's ability based on whether the answer is correct or incorrect (NCSBN, 2019b). The computer then searches the item bank for an item with a 50% chance of being answered correctly by the candidate. This is an efficient means of testing, avoiding questions that do not contribute to determining a candidate's level of nursing competence.

The standard for passing the NCLEX is criterion referenced. The standard is set by the NCSBN based on an established protocol and is used as the basis for determining whether the candidate has passed or failed the examination. The NCLEX-RN can range from 75 to 265 items, with 15 of those being pretest items that are not scored. After candidates answer the minimum number of items, the testing stops when the candidate's ability is above or below the standard for passing, with 95% certainty (NCSBN, 2019b). Because the NCLEX is an adaptive test, candidates complete different numbers of items, and therefore the test takes varying amounts of time. If a candidate's ability has not been determined by the time the maximum number of items has been presented or when the time limit has been reached, the examination then stops.

All RN candidates must answer a minimum number of 75 items. The maximum number they can answer is 265 within a time limit of 6 hours (NCSBN, 2019b). On the NCLEX-PN, PN and VN candidates must answer a minimum of 85 items. The maximum number of items they can answer is 205, during the 5-hour testing period allowed (NCSBN, 2016).

■ Preparation of Items at Varied Cognitive Levels

When courses have higher level outcomes, tests in those courses need to measure learning at the applying and analyzing levels rather than at remembering and understanding. This principle was discussed in earlier chapters. Items at higher levels of cognitive complexity are more difficult and time-consuming to develop, but they provide a way of evaluating ability to apply knowledge to new situations and to engage in analytical thinking. The majority of items on the NCLEX are written at higher levels of cognitive ability, requiring application of knowledge and analytical thinking.

Students are at a disadvantage if they encounter only test items that ask them to recall facts as they progress through a nursing program. Low-level items assess how well students memorize specific information, not whether they can use that knowledge to analyze clinical situations and arrive at the best decisions possible for those situations. Students need experience answering questions at the application and analysis levels before they take the NCLEX. More important, if course outcomes are at higher levels of cognitive complexity, then tests and other methods need to assess learning at

those levels. In graduate nursing programs, test items should be developed at higher cognitive levels to assess students' ability to problem solve and think through complex situations to prepare them for certification examinations they might take as graduates.

When developing a new test, a blueprint is important in planning the number of items at each cognitive level for the content areas to be assessed. By using a blueprint, teachers can avoid writing too many items that require only recall of information. For existing tests that were not developed using a blueprint, teachers can code items using Bloom's taxonomy or the updated taxonomy of the cognitive domain and then decide whether more higher level items should be added.

Remembering (Knowledge)

In developing items at varying cognitive levels, it is important to remember the learning outcome intended at each of these levels. Questions at the remembering or knowledge level deal with facts, principles, and other specific information that is memorized and then recalled to answer the item. An example of a multiple-choice item at the remembering level follows:

Your patient is taking pseudoephedrine for his stuffy nose. Which of the following side effects is common among patients using this medication?

- a. Diarrhea
- b. Dyspnea
- c. Hallucinations
- d. Restlessness¹

Understanding (Comprehension)

At this level, items assess understanding of concepts and ability to explain them. These questions are written at a higher level than remembering facts, but they do not assess use of information in a new context. An example of an item at the understanding level is:

A 30-year-old sexually active woman has amenorrhea and bloody vaginal spotting. The nurse practitioner finds that the patient's left adnexa is tender and she has cervical motion tenderness. Which test should the nurse practitioner order first?

- a. Complete blood count (CBC) with white cell differentials
- b. Flat plate of the abdomen
- c. Pelvic ultrasound
- d. Urine pregnancy test¹

¹Correct answer.

Applying (Application)

At the applying level, students apply concepts and other types of knowledge as a basis for responding to the item. At this level, test questions measure *use* of knowledge in new or unique situations. One method for developing items at this level is to prepare stems that have information that students did not encounter in their learning about the content. The stem might present patient data, problems, or interventions different from the ones discussed in class or in the readings. If examples related to nursing care of adults, items might test the ability to use those concepts when the patient is an adolescent or has multiple coexisting conditions. An example of an item at the applying level is as follows:

A mother tells you that she is worried about her 4-year-old daughter's development because her daughter seems to be "behind." You complete a developmental assessment. Which of the following behaviors suggests the need for further developmental testing?

- a. Cannot follow five commands in a row
- b. Has difficulty holding a crayon between thumb and forefinger¹
- c. Is unable to balance on each foot for 6 seconds
- d. Keeps making mistakes when asked about the day of the week

Analyzing (Analysis)

Questions at the analyzing level are the most difficult to construct. They require analysis of a clinical or other situation to identify critical elements and relationships among them. Items should provide a new situation for students to analyze, not one encountered previously for which the student might recall the analysis. Many of these items require learners to solve a problem and make a decision about priorities or the best approach to take among the options. Or items might ask students to identify the most immediate course of action to meet patient needs or manage the clinical situation.

The difference between applying and analyzing items is not always readily apparent. Items at the analyzing level, though, should ask students to identify relevant data, critical elements in the scenario, and their interrelationships. In analysis-level items, students should distinguish between significant and nonsignificant information and select the best approach or priority among those cited in the alternatives. An example of an item written at the analysis level is as follows:

You receive a report on the following patients at the beginning of your evening shift at 3 p.m. Which patient should you assess first?

- a. An 82-year-old with pneumonia who seems confused at times¹
- b. A 76-year-old patient with cancer with 300 mL remaining of an intravenous infusion
- c. A 40-year-old who had an emergency appendectomy 8 hours ago
- d. An 18-year-old with chest tubes for treatment of a pneumothorax following an accident

■ Preparation of Items Within the Framework of Clinical Practice

One of the processes integrated into the NCLEX test plans is the nursing process. This is also a framework taught in many nursing programs. If not presented as the nursing process, most clinical courses address, in some form, assessment, data analysis, problems or diagnoses, interventions, and evaluation. These areas provide another useful framework for developing test questions. Items can examine assessment of patients with varied needs and health problems, analysis of data, identification of problems, selection of evidence-based interventions and treatments, and evaluation of the outcomes of care.

Current practices suggest that many test items focus on scientific rationale, principles underlying patient care, and selection of interventions. Fewer items are developed on collecting and analyzing data, determining patient problems, setting priorities and realistic goals of care, and evaluating the effectiveness of interventions and outcomes. Developing items based on clinical scenarios in the stems provides an opportunity to examine these outcomes of learning. These items also facilitate testing at a higher cognitive level because they are written in relation to specific scenarios in the stems, requiring students to apply their knowledge to the clinical situation. Items may stand alone, or a series of items may be developed related to one clinical scenario. In the latter format, the teacher has an option of adding data to the situation and creating an unfolding case, which was discussed in Chapter 7, Assessment of Higher Level Learning.

The stems in Exhibit 8.1 can be used to develop test items that evaluate students' ability to assess patients and identify data to collect in a clinical situation, analyze data, determine priority problems, select the correct or best intervention, and evaluate effectiveness of care. They can be integrated in a stem with a clinical scenario that relates to the content of the course being tested or used as stems for a multiple-choice, multiple-response, or other item type not related to a specific clinical situation.

EXHIBIT 8.1**EXAMPLES OF STEMS FOR CLINICAL PRACTICE ITEMS****ASSESSMENT**

The nurse should collect which of the following data?

Which of the following information should be collected as a priority in the assessment?

Which data should be collected first?

Which questions should the nurse ask (the patient, the family, others) in the assessment?

Your patient develops (symptoms). What data should the nurse collect now?

What additional data are needed to establish the patient's problems?

Which resources should be used to collect the data?

Which of the following information is a priority to report in an SBAR (situation–background–assessment–recommendation) communication to the (physician, nurse, other provider)?

ANALYSIS

These data support the (diagnosis, problem) of _____.

Which (diagnosis, problem) is most appropriate for this patient?

The priority nursing diagnosis is _____.

The priority problem of this (patient, family, community) is _____.

A patient with (a diagnosis of, symptoms of) is at risk for developing which of the following complications?

PLANNING

Which outcomes are most important for a patient with a (problem of)?

What are the priority outcomes for a patient receiving (treatment)?

Which nursing measures should be included in the plan of care for a patient with (problem, surgery, treatment, diagnostic test)?

Which of the following nursing interventions would be most effective for a patient with (diagnosis of, problem of, symptoms of)?

The nurse is teaching a patient who is (years old). Which teaching strategy would be most appropriate?

Which intervention is most likely to be effective in managing (symptoms of)?

IMPLEMENTATION

Which of the following actions should be implemented immediately?

Nursing interventions for this patient include _____.

Following this (procedure, surgery, treatment, test), which nursing measures should be implemented?

Which of these nursing interventions is a priority for a patient with (problem)?

What evidence supports (nursing intervention)?

A patient with (a diagnosis of) complains of (symptoms). What should the nurse do first?

(continued)

EXHIBIT 8.1**EXAMPLES OF STEMS FOR CLINICAL PRACTICE ITEMS (*continued*)**

Which explanation should the nurse use when teaching a patient (with a diagnosis of, prior to procedure, surgery, treatment, test)?

Which of the following instructions should be given to the (patient, family, caregiver, nurse) at discharge?

Which of the following situations (incidents) should be reported immediately?

EVALUATION

Which of these responses indicates that the (intervention, medication, treatment) is effective?

A patient is taking (medication) for (diagnosis, problem). Which of these data indicate a side effect of the medication?

Which response by the patient indicates improvement?

Which of the following observations indicates that the (patient, caregiver) knows how to (perform the procedure, give the treatment, follow the protocol)?

Which statement by the (patient, caregiver) indicates the need for further teaching?

Teachers can select a stem from Exhibit 8.1 and add content from their own course, providing an easy way of writing items within the framework of clinical practice or the nursing process and evaluating exemplars of concepts taught in the course or curriculum. Sample items are as follows:

Assessment

An 8-year-old boy is brought to the emergency department by his mother after falling off his bike and hitting his head. Which of the following data is most important to collect in the initial assessment?

- a. Blood pressure
- b. Level of consciousness
- c. Pupillary response
- d. Respiratory status¹

Analysis

The nurse practitioner is admitting a patient who complains of fatigue and myalgia, and has a rash across the bridge of the nose and cheeks. The practitioner finds a few ulcers in the patient's mouth. Prior laboratory tests included a positive C-reactive protein. These findings support a likely diagnosis of

- a. Fibromyalgia
- b. Rheumatoid arthritis
- c. Scleroderma
- d. Systemic lupus erythematosus¹

Planning

Your patient is being discharged after a sickle cell crisis. Which of the following measures should be included in your teaching plan for this patient? Select all that apply.

- ☐ 1. Avoid warm temperatures inside and outdoors.
- ☐ 2. Do not use nonsteroidal anti-inflammatory drugs for pain.
- ☒ 3. Drink at least eight glasses of water a day.
- ☒ 4. Eat plenty of grains, fruits, and green leafy vegetables.
- ☒ 5. Get a vaccination for pneumonia.
- ☐ 6. Keep cold packs handy for joint pain.

Implementation

Your patient is in active labor with contractions every 3 minutes lasting about 1 minute. She appears to have a seizure. Which of the following interventions is the top priority?

- a. Assess her breathing pattern¹
- b. Attach an external fetal monitor
- c. Call the physician
- d. Prepare for a cesarean delivery

Evaluation

A male adult patient was discharged following a below-the-knee amputation. You are making the first home health visit after his discharge. Which of the following statements by the patient indicates that he needs further instruction?

- a. "I know to take my temperature if I get chills again like in the hospital."
- b. "I won't exert myself around the house until I see the doctor."
- c. "The nurse said to take more insulin when I start to eat more."¹
- d. "The social worker mentioned a support group. Maybe I should call about it."

■ Preparation of Students for the NCLEX Examinations

A number of studies have been done over the years to identify predictors of success on the NCLEX-RN. Some factors related to performance on the NCLEX-RN are Scholastic Aptitude Test (SAT), American College Testing (ACT), and other preadmission test scores (Grossbach & Kuncel, 2011; Robert, 2018; Wambuguh, Eckfield, & Van Hofwegen, 2016); scores on NCLEX readiness tests (Brodersen & Mills, 2014; Brussow & Dunham, 2018; Penprase & Harris, 2013; Schooley & Kuhn, 2013; Sosa & Sethares, 2015); standardized assessment scores for specific content areas, for example, fundamentals, pharmacology, and medical-surgical nursing (Chen & Bennett, 2016; Emory, 2013, 2019; McCarthy, Harris, & Tracz, 2014; Twidwell & Records, 2017; Yeom, 2013); grades in nursing courses and graduation grade point average (Grossbach & Kuncel, 2011; Kaddoura, Flint, Van Dyke, Yang, & Chiang, 2017; Romeo, 2013); being a transfer student (Simon, McGinniss, & Krauss, 2013); grades in science courses (Elder, Jacobs, & Fast, 2015; Wambuguh et al., 2016); and student characteristics, such as students who speak English as a second or an additional language (Kaddoura et al., 2017), and emotional intelligence (Rode & Brown, 2019). Academic achievement, in terms of nursing course grades and overall grade point average, has been found across studies as predictive of student performance on the NCLEX-RN. In a meta-analysis of 31 independent samples with 7,159 participants, admissions test scores (SAT and ACT) and grades earned in nursing programs, especially grades in the second year, were the best predictors of performance on the NCLEX (Grossbach & Kuncel, 2011).

A second area of the literature on the NCLEX-RN focuses on methods of preparing students to pass the examination. Many schools use standardized examinations designed to predict student performance on the NCLEX-RN and determine students' readiness for taking the examination. By analyzing the results of standardized tests for NCLEX readiness, faculty members and students can work together to design individual plans for remediation so that students will be more likely to experience first-time success on the licensure examination.

Some of these approaches include curriculum review to identify areas needing improvement, using readiness tests combined with remedial learning activities to better prepare students for the NCLEX, student self-assessment of content areas needing improvement, instruction for content mastery, test-taking tips, managing test anxiety, cooperative study groups, courses that guide formal NCLEX-RN preparation, and careful planning for the day of testing. Comprehensive mentoring and coaching programs provide the support students and graduates need as they prepare for taking the NCLEX (Havrilla, Zbegner, & Victor, 2018; McKelvey et al., 2018; Schlairet & Rubenstein, 2019). Experience with test items that are similar to the NCLEX prepares students for the types of items they will encounter on the licensing examination. In addition to these item formats, students also need experience in taking practice tests.

■ Summary

The chapter summarized the NCLEX test plans and their implications for nurse educators. One of the principles emphasized was the need to prepare items at different cognitive levels as indicated by the outcomes of the course. Items at the remembering or knowledge level assess how well students memorized facts and specific information; they do not, however, provide an indication of whether students can use that information in practice or can engage in higher level thinking. To assess those higher level outcomes, items must be written at the applying or analyzing levels or evaluated by methods other than tests. It is worthwhile for faculty members to develop a test blueprint that specifies the number of items to be developed at each cognitive level for content areas in the course. By using a blueprint, teachers can avoid writing too many lower level items on an examination.

As students progress through a nursing program, they develop knowledge and skills to assess patients, analyze data, identify patient needs and problems, set priorities for care, select appropriate and evidence-based interventions, and evaluate the outcomes of care. Testing within the framework of clinical practice or the nursing process provides an opportunity to assess those learning outcomes. Items may be written about data to collect in the particular clinical scenario, possible problems, approaches to use, priorities of care, decisions to be made, varying judgments possible in a scenario, and other questions that examine students' thinking and clinical judgment as related to the situation described in the stem of the item. This format of testing also provides experience for students in answering the types of items encountered on licensure and certification examinations.

References

- Brodersen, L. D., & Mills, A. C. (2014). A comparison of two nursing program exit exams that predict first-time NCLEX-RN outcome. *Computers, Informatics, Nursing*, 32, 404–412. doi:10.1097/CIN.0000000000000081
- Brussow, J. A., & Dunham, M. (2018). Students' midprogram content area performance as a predictor of end-of-program NCLEX readiness. *Nurse Educator*, 43, 238–241. doi:10.1097/nne.0000000000000499
- Chen, H. C., & Bennett, S. (2016). Decision-tree analysis for predicting first-time pass/fail rates for the NCLEX-RN(R) in associate degree nursing students. *Journal of Nursing Education*, 55, 454–457. doi:10.3928/01484834-20160715-06
- Elder, B. L., Jacobs, P., & Fast, Y. J. (2015). Identification and support of at-risk students using a case management model. *Journal of Professional Nursing*, 31, 247–253. doi:10.1016/j.profnurs.2014.10.003
- Emory, J. (2013). Standardized mastery content assessments for predicting NCLEX-RN outcomes. *Nurse Educator*, 38, 66–70. doi:10.1097/NNE.0b013e3182829c94
- Emory, J. (2019). Exploring NCLEX failures and standardized assessments. *Nurse Educator*, 44, 142–146. doi:10.1097/nne.0000000000000601
- Grossbach, A., & Kuncel, N. R. (2011). The predictive validity of nursing admission measures for performance on the National Council Licensure Examination: A meta-analysis. *Journal of Professional Nursing*, 27, 124–128. doi:10.1016/j.profnurs.2010.09.010
- Havrilla, E., Zbegner, D., & Victor, J. (2018). Exploring predictors of NCLEX-RN success: One school's search for excellence. *Journal of Nursing Education*, 57, 554–556. doi:10.3928/01484834-20180815-08
- Kaddoura, M. A., Flint, E. P., Van Dyke, O., Yang, Q., & Chiang, L. C. (2017). Academic and demographic predictors of NCLEX-RN pass rates in first- and second-degree accelerated BSN programs. *Journal of Professional Nursing*, 33, 229–240. doi:10.1016/j.profnurs.2016.09.005
- McCarthy, M. A., Harris, D., & Tracz, S. M. (2014). Academic and nursing aptitude and the NCLEX-RN in baccalaureate programs. *Journal of Nursing Education*, 53, 151–159.
- McKelvey, M. M., Langevin, K. M., Konieczny, L., Espelin, J. M., Peer, N., Christensen, S., & Thomas, C. (2018). Nursing faculty coaches: Uncovering a hidden resource for NCLEX-RN success. *Creative Nursing*, 24, 225–230. doi:10.1891/1078-4535.24.4.225
- National Council of State Boards of Nursing. (2016). *NCLEX-PN® examination: Test plan for the National Council Licensure Examination for Licensed Practical/Vocational Nurses*. Chicago, IL: Author. Retrieved from https://www.ncsbn.org/2017_PN_TestPlan.pdf
- National Council of State Boards of Nursing. (2018). *2017 RN practice analysis: Linking the NCLEX-RN® examination to practice: U.S. and Canada*. Chicago, IL: Author. Retrieved from https://www.ncsbn.org/17_RN_US_Canada_Practice_Analysis.pdf
- National Council of State Boards of Nursing. (2019a). *Next generation NCLEX project*. Chicago, IL: Author. Retrieved from <https://www.ncsbn.org/next-generation-nclex.htm>
- National Council of State Boards of Nursing. (2019b). *NCLEX-RN® examination: Test plan for the National Council Licensure Examination for Registered Nurses*. Chicago, IL: Author. Retrieved from https://www.ncsbn.org/2019_RN_TestPlan-English.pdf
- National Council of State Boards of Nursing. (2019c). *2018 LPN/VN practice analysis: Linking the NCLEX-PN® examination to practice*. Chicago, IL: Author. Retrieved from https://www.ncsbn.org/LPN_Practice_Analysis_FINAL.pdf

- Penprase, B. B., & Harris, M. A. (2013). Accelerated second-degree nursing students: Predictors of graduation and NCLEX-RN first-time pass rates. *Nurse Educator*, 38, 26–29. doi:10.1097/NNE.0b013e318276df16
- Robert, N. (2018). Predictors of program completion and NCLEX-RN success in an associate degree nursing program. *Nursing Education Perspectives*, 39, 38–39. doi:10.1097/01.Nep.0000000000000237
- Rode, J., & Brown, K. (2019). Emotional intelligence relates to NCLEX and standardized readiness test: A pilot study. *Nurse Educator*, 44, 154–158. doi:10.1097/nne.0000000000000565
- Romeo, E. M. (2013). The predictive ability of critical thinking, nursing GPA, and SAT scores on first-time NCLEX-RN performance. *Nursing Education Perspectives*, 34, 248–253.
- Schlairet, M. C., & Rubenstein, C. (2019). Senior NCLEX-RN coaching model: Development and implementation. *Nurse Educator*, 44, 250–254. doi:10.1097/nne.0000000000000644
- Schooley, A., & Kuhn, J. R. (2013). Early indicators of NCLEX-RN performance. *Journal of Nursing Education*, 52, 539–542. doi:10.3928/01484834-20130819-08
- Simon, E. B., McGinniss, S. P., & Krauss, B. J. (2013). Predictor variables for NCLEX-RN readiness exam performance. *Nursing Education Perspectives*, 34, 18–24.
- Sosa, M. E., & Sethares, K. A. (2015). An integrative review of the use and outcomes of HESI testing in baccalaureate nursing programs. *Nursing Education Perspectives*, 36, 237–243.
- Twidwell, J. E., & Records, K. (2017). An integrative review on standardized exams as a predictive admission criterion for RN programs. *International Journal of Nursing Education Scholarship*, 14, pii. doi:10.1515/ijnes-2016-0040
- Wambuguh, O., Eckfield, M., & Van Hofwegen, L. (2016). Examining the importance of admissions criteria in predicting nursing program success. *International Journal of Nursing Education Scholarship*, 13(1), pii. doi:10.1515/ijnes-2015-0088
- Woo, A., & Dragan, M. (2012). Ensuring validity of NCLEX® with differential item functioning analysis. *Journal of Nursing Regulation*, 2(4), 29–31.
- Yeom, Y. J. (2013). An investigation of predictors of NCLEX-RN outcomes among nursing content standardized tests. *Nurse Education Today*, 33, 1523–1528. doi:10.1016/j.nedt.2013.04.004

ASSESSMENT OF WRITTEN ASSIGNMENTS

In most nursing courses, students complete some type of written assignment. With these assignments, students can develop their critical thinking skills, gain experience with different types of writing, and achieve other outcomes specific to a course. Written assignments with feedback from the teacher help students develop their writing ability, which is an important outcome in any nursing program from the beginning level through graduate study. This chapter focuses on developing and assessing written assignments for nursing courses.

■ Purposes of Written Assignments

Written assignments are a major instructional and assessment method in nursing courses. They can be used to achieve many learning outcomes, but need to be carefully selected and designed with consideration of the instructional goals. With written assignments students can (a) critique and synthesize the literature and report on their findings; (b) search for, critique, and integrate evidence for nursing practice; (c) analyze concepts and theories and apply them to clinical situations; (d) improve their problem-solving and higher level thinking skills; (e) gain experience in formulating their ideas and communicating them in a clear and coherent way to others; and (f) develop writing skills. Many of the written assignments in clinical courses assist students in thinking through their plan of care and identifying areas in which they need further instruction. Some assignments, such as reflective journals, also encourage students to examine their own feelings, beliefs, and values and to reflect on their learning in a course.

Not all written assignments achieve each of these purposes, and the teacher plans the assignment based on the intended goals of learning. Assignments should meet *specific* objectives of a course and should not be included only for the purpose of

having a written assignment as a course requirement. Instead, they should be carefully selected to help students improve their writing skills and achieve course outcomes.

Because writing is a developmental process that improves with practice, writing assignments should build on one another throughout a course and throughout the entire nursing program. Writing a sequence of papers across courses encourages the improvement of writing more effectively than having students complete a different type of paper in each course. This planning also eliminates excessive repetition of assignments in the program. Along the same lines, faculty members should decide the number of written assignments needed by students to achieve the outcomes of a course or clinical practice experience. In some clinical nursing courses, students complete the same assignments repeatedly throughout a course, leading to their frustration with the “paperwork” in the course. How many times do students need to submit a written assessment of a patient? Written assignments are time-consuming for students to prepare and teachers to read and respond to. Thus, such assignments should be carefully selected to meet course goals and should benefit the students in terms of their learning.

Writing in the Discipline and Writing-to-Learn Activities

Writing assignments in a nursing course in which students receive feedback on their writing guides students in learning how to write clearly for varied audiences. The ability to communicate ideas in writing is an essential outcome of a nursing program at all educational levels. The dissemination of new ideas and innovations, outcomes of clinical projects, and findings of research studies and quality improvement projects requires skill in writing. This skill can be developed through formal papers in a nursing course, such as term papers, in which students receive feedback on their writing; this is often referred to as *writing in the discipline* (Writing Across the Curriculum [WAC] Clearinghouse, 2019a). With formal papers that students prepare in a course, the teacher can give feedback on the content of the paper and also on the quality of the writing (Oermann, 2013; Oermann et al., 2015). The experience of writing a paper and being guided on its development, combined with feedback from the teacher, allows students to develop their skills in thinking through the topic of the paper and how best to communicate their ideas in writing. These assignments are typically completed over a period of time, and both the content and writing skill are assessed. Formal papers also provide an opportunity to learn a reference style, such as the *Publication Manual of the American Psychological Association* (APA), although this is only one aspect of the assessment (APA, 2010). The goal with written assignments such as formal papers is to learn to write effectively and communicate ideas clearly, not only how to use APA or another writing style.

Formal papers can be divided into smaller writing assignments and sequenced progressively through a course. This makes completion of the paper more manageable

for students, allows the teacher to assess and provide feedback on each part of the paper, and encourages students to use that feedback for revisions as they are preparing the longer paper. Shorter and carefully planned assignments can be integrated across courses to provide practice in writing. This avoids students preparing a formal paper in one course and having no other written assignments for which they receive feedback on writing in their other courses. Luthy, Peterson, Lassetter, and Callister (2009) suggested that smaller assignments that build on one another are less daunting for students. For example, students might be asked to prepare a paper on a potential safety issue in the clinical setting using the National Patient Safety Goals (The Joint Commission, 2019). The first assignment might be the description of the clinical setting and patient population, an issue they identified with supporting data, and a relevant safety goal. The second assignment might be a literature review related to the safety goal such as the need to communicate important test results to the right person on time, why this is important, and initiatives to improve staff communication. The third assignment might be a description of the initiative they selected for implementation on their unit, their rationale based on the literature and their analysis of the clinical setting, and a plan for implementation and evaluation of outcomes. With each of these written assignments, the teacher can provide feedback, followed by student revision of both the content and writing.

Other types of writing assignments in a nursing course, such as reflective journals and in-class writing activities, guide students in reflecting on their experiences or learning course content but do not promote development of writing ability. These are writing-to-learn activities: They are typically short and informal, and may be impromptu, but they help students explore their understanding of content and think through key concepts presented in class (Halim, Finkenstaedt-Quinn, Olsen, Gere, & Shultz, 2018; Oermann, 2013; WAC Clearinghouse, 2019b). For example, if students are unclear about a topic presented in class, they can be asked to write down their questions or summarize in their own words key points that they learned in class. The outcome of this type of writing activity is *learning*, not improving writing skill. The assessment would be formative with feedback on the content, not the writing.

Drafts and Rewrites

Formal papers enable the teacher to assess students' ability to present, organize, and express ideas effectively in writing. Through these written assignments, students develop an understanding of the content they are writing about, and they learn how to communicate their ideas in writing. To improve their writing abilities, though, students need to complete drafts of papers on which they receive feedback from the teacher.

Drafts and rewrites of papers are essential if the goal is to develop skill in writing (Oermann, 2013; Oermann et al., 2015). Teachers should critique papers for quality

of the content; organization; process of developing ideas and arguments; and writing style such as clarity of expression, sentence structure, punctuation, grammar, spelling, length of the paper, and accuracy and format of the references (Oermann & Hays, 2019). This critique should be accompanied by feedback on how to improve writing. Students need specific suggestions about revisions, not general statements such as “*writing is unclear*.” Instead, the teacher should identify the problem with the writing and give suggestions as to how to improve it, for example, “*Introductory sentence does not relate to the content in the paragraph. Replace it with a sentence that incorporates the three nursing measures you discuss in the paragraph.*” Drafts combined with feedback from the teacher are intended to improve students’ writing skills. Because they are used for this purpose, they should not be graded.

Providing feedback on writing is time-consuming for teachers. Another method that can be used is for students to critique each other’s writing in small groups or pairs. Peers can provide valuable feedback on content, organization, how the ideas are developed, and whether the writing is clear. Although they may not identify errors in grammar and sentence structure, they often can find problems with errors in content and clarity of writing. Students can post sections of their papers or questions about writing online in a discussion forum or prepare papers using collaborative tools such as Google Docs; peers and the teacher can provide feedback and answer the questions. This collaboration benefits not only the individual student writing the paper but also the group as a whole. Peers also can assess writing in small-group activities in the classroom, online, and in postclinical conference if the writing assignment deals with clinical practice. Small-group critique provides a basis for subsequent revisions.

■ Types of Written Assignments

Many types of writing assignments are appropriate for assessment in nursing education. Some of these assignments provide information on how well students have learned the content but do not necessarily improve their writing skill. For example, structured assignments that involve short sentences and phrases, such as nursing care plans and write-ups of assessments and physical examinations, do not foster development of writing skills nor do they provide sufficient data for assessing writing.

Other assignments such as formal papers can be used for assessing students’ understanding as well as writing ability. Therefore, not all written assignments provide data for assessing writing skill, and again the teacher needs to be clear about the outcomes to be evaluated with the assignment. Many written assignments can be used in nursing courses. These include:

- Term paper
- Research paper and development of research protocol

- Literature reviews and integrative reviews
- Evidence-based practice paper in which students critique and synthesize the evidence and report on its use in clinical practice
- Paper analyzing concepts and their application to clinical practice
- Paper comparing different interventions with their underlying evidence base
- Paper on how the content they learned in class and read about in their textbook and articles compares with their experiences in the clinical setting and how it applies to patient care
- Critical analysis papers in which students analyze issues, compare different options, and develop arguments for a position
- Case study analysis with written rationale
- Reflective journals and writing assignments

For clinical courses, written assignments that accompany the clinical practicum are valuable for learning to use evidence-based resources and developing clinical judgment skills. They also provide a strategy for students to analyze ethical issues in the clinical setting and reflect on their personal experiences with patients and staff. Writing assignments such as reflective journals bridge the gap between classroom learning and clinical practice, encourage students to think about practice decisions and evaluate choices, promote self-awareness and professional growth, and provide valuable feedback for faculty in identifying students having difficulty with clinical judgment and other learning needs (Bussard, 2015; Dahl & Eriksen, 2016; Lasater, 2011). Short papers in clinical courses are useful in focusing an assignment on a particular learning outcome and making it easier for teachers to give prompt feedback to students (Oermann, Shellenbarger, & Gaberson, 2018). For example, students might write a one-page paper on an alternate intervention for a patient with evidence for its use, or prepare a short paper on an issue encountered in clinical practice and an alternate approach that could have been used.

Written assignments for clinical learning include:

- Concept map, a graphic arrangement of key concepts related to a patient's care, which includes a written description of the meaning of the interrelationships
- Analysis of a clinical experience, the care given by the student, and alternative approaches that could have been used
- Paper that examines how readings apply to care of the patient
- Short paper related to clinical practice
- Teaching plan
- Nursing care plan
- Analysis of interactions with individuals and groups in the clinical setting

- Report of observations made in clinical settings
- Reflective journal and other reflective writing activities
- e-Portfolio, a collection of projects and materials that demonstrate student learning in clinical practice

In-Class and Small-Group Writing Activities

Not all written assignments need to be prepared by students individually as out-of-class work that is assessed by the teacher. In-class writing assignments provide practice in expressing ideas and an opportunity for faculty and peers to give feedback on writing. For example, students can write a summary of the key content areas presented in a face-to-face class or online. In a concept-based curriculum, they might identify exemplars of a concept that were not included in the class. Students can list one or two questions about the content and give the questions to other students to answer in writing or to post in a discussion forum. The teacher can pose a question about how the content could be applied in a different context and ask students to write a response to the question. In a face-to-face class, several students can read their responses aloud, and the teacher can collect all written responses for later analysis. In an online course, students can post their individual responses for critique by other students. An activity such as this one assists students in organizing their thoughts before responding to questions raised by the teacher and others. Another option is for students to write a few paragraphs about how the content compares with their readings: What new learning did they gain from the class that was not in their readings?

As another writing activity, the teacher can give students short case studies related to the content being learned in the course. In small groups or individually, students analyze these cases, identify possible problems, and develop plans of care, and then report in a few paragraphs the results of their analysis and rationale for their plan. They also can describe in writing how the case is similar to or differs from what they learned in class or from their readings.

These short written activities are valuable at the end of a class to summarize the new content and actively involve students in learning. With any of these activities, students can “pass their writing” to peers whose task is to critique both content and writing, adding their own thoughts about the topic and assessing the writing. The teacher also can review the written work to provide feedback.

Students can work in pairs or small groups for writing assignments. For example, a small group of students can write an editorial or a letter to the editor; develop a protocol for patient care based on the content presented in the lecture and readings for class; and review, critique, and summarize evidence that relates to patient care.

At the graduate level, students can prepare a manuscript or work through the steps in writing for publication beginning with an outline, preparing a draft, and revising the draft for a final product. These assignments among others encourage acquisition of content and development of skill in writing; they also provide experience in group writing, learning about its benefits and pitfalls.

Writing Activities for Postclinical Conferences

In postclinical conferences, students can work in pairs or in small groups to critically analyze a clinical situation, decide on alternate interventions that might be used, and then write a short paper about their discussion. They can write about their own clinical activities and document the care they provided during that clinical experience. “Pass the writing” assignments work well in clinical conferences because they encourage peers to critically analyze the content, adding their own perspectives, and to identify how writing can be improved. These assignments also actively involve students in learning, which is important during a tiring clinical practicum. Group writing exercises are effective in postclinical conferences as long as the groups are small and the exercises are carefully focused.

■ Assessing Written Assignments

Papers and other types of written assignments should be assessed using predetermined criteria that address quality of content; organization of ideas; and the process of arriving at decisions and, depending on the assignment, developing an argument. Writing style should also be considered. General criteria for this purpose, which can be adapted for most written assignments, are found in Exhibit 9.1.

Scoring rubrics work well for assessing papers. A rubric is a scoring guide used for the assessment of performance. Rubrics outline the criteria to meet in the paper, or describe the characteristics of the paper and the points allotted for its assessment. Rubrics lead to fairer, more transparent, and more consistent scoring of papers (Fulbright, 2018; Minnich et al., 2018). The points assigned to each criterion or characteristic in the rubric should reflect its importance in the paper. A description and examples of holistic and analytic scoring rubrics were provided in Chapter 6, Short-Answer (Fill-in-the-Blank) and Essay. Rubrics should be given to students before they begin writing so they are clear about how the paper will be assessed. In this way, the rubric can be viewed as an instructional guide and assessment tool (Brookhart & Nitko, 2019). An example of a rubric for scoring papers and other written assignments, based on the general criteria outlined in Exhibit 9.1, is shown in Table 9.1.

EXHIBIT 9.1**CRITERIA FOR ASSESSING PAPERS AND OTHER WRITTEN ASSIGNMENTS****Content**

- Content is relevant.
- Content is accurate.
- Significant concepts and theories are presented.
- Concepts and theories are used appropriately for analysis.
- Content is comprehensive.
- Content reflects current research and evidence.
- Hypotheses, conclusions, and decisions are supported.

Organization

- Content is organized logically.
- Ideas are presented in logical sequence.
- Paragraph structure is appropriate.
- Headings are used appropriately to indicate new content areas.

Process

- Process used to arrive at approaches, decisions, judgments, and so forth, is adequate.
- Consequences of decisions are considered and weighed.
- Sound rationale is provided based on theory and research as appropriate.
- For papers analyzing issues, rationale supports the position taken.
- Multiple perspectives and new approaches are considered.

Writing Style

- Ideas are described clearly.
- Sentence structure is clear.
- There are no grammatical errors.
- There are no spelling errors.
- Appropriate punctuation is used.
- Writing does not reveal bias related to gender, sexual orientation, racial or ethnic identity, or disabilities.
- Length of paper is consistent with requirements.
- References are cited appropriately throughout the paper.
- References are cited accurately according to the required format.

Source: Oermann, M. H., Shellenbarger, T., & Gaberson, K. B. (2018). *Clinical teaching strategies in nursing* (5th ed.). New York, NY: Springer Publishing Company. Copyright 2018 by Springer Publishing Company. Reprinted with permission.

TABLE 9.1 Sample Scoring Rubric for Term Papers and Other Written Assignments

CONTENT		
Content relevant to purpose of paper, comprehensive and in depth	Content relevant to purpose of paper	Some content not relevant to purpose of paper, lacks depth
10 9 8	7 6 5 4	3 2 1
Content accurate	Most of content accurate	Major errors in content
10 9 8	7 6 5 4	3 2 1
Sound background developed from concepts, theories, and literature	Background relevant to topic but limited development	Background not developed, limited support for ideas
20–15	14–7	6–1
Current research synthesized and integrated effectively in paper	Relevant research summarized in paper	Limited research in paper, not used to support ideas
10 9 8	7 6 5 4	3 2 1
ORGANIZATION		
Purpose of paper/thesis well developed and clearly stated	Purpose/thesis apparent but not developed sufficiently	Purpose/thesis poorly developed, not clear
5	4 3 2	1
Ideas well organized and logically presented, organization supports arguments and development of ideas	Clear organization of main points and ideas	Poorly organized, ideas not developed adequately in paper
10 9 8	7 6 5 4	3 2 1
Thorough discussion of ideas, includes multiple perspectives and new approaches	Adequate discussion of ideas, some alternate perspectives considered	Discussion not thorough, lacks detail, no alternate perspectives considered
10 9 8	7 6 5 4	3 2 1
Effective conclusion and integration of ideas in summary	Adequate conclusion, summary of main ideas	Poor conclusion, no integration of ideas
5	4 3 2	1
WRITING STYLE AND FORMAT		
Sentence structure clear, smooth transitions, correct grammar and punctuation, no spelling errors	Adequate sentence structure and transitions; few grammar, punctuation, and spelling errors	Poor sentence structure and transitions; errors in grammar, punctuation, and spelling
10 9 8	7 6 5 4	3 2 1

(continued)

TABLE 9.1 Sample Scoring Rubric for Term Papers and Other Written Assignments (*continued*)

Professional appearance of paper, all parts included, length consistent with requirements	Paper legible, some parts missing or too short/ too long considering requirements	Unprofessional appearance, missing sections, paper too short/ too long considering requirements
5	4 3 2	1
References used appropriately in paper, references current, no errors in references, correct use of APA style for references	References used appropriately in paper but limited, most references current, some citations or references with errors and/or some errors in APA style for references	Few references and limited breadth, old references (not classic), errors in references, errors in APA style for references
5	4 3 2	1
Total Points _____ (sum points for total score)		

APA, American Psychological Association.

Consistent with other evaluation methods, written assignments may be assessed either formatively (not graded) or summatively (graded). With formative evaluation, the intent is to give feedback on the quality of the content and writing so that students can further develop their writing ability. Feedback is of value only if given promptly and with enough detail for students to understand how they can improve their writing. With some assignments, such as reflective journals, only formative evaluation may be appropriate.

Many nursing faculty members are concerned about the amount of time spent giving feedback on students' technical writing errors, such as grammatical, punctuation, and spelling errors. If teachers focus entirely on assessing the quality of content of written assignments, students will not understand how their technical writing skills affect their ability to communicate relevant and important information. There is a difference between giving feedback on the quality of technical writing skills and actually correcting errors for students. One method for avoiding the latter approach on a graded assignment is to signify technical writing errors with a particular symbol such as a checkmark, or more specific, by identifying the type of error, such as "spelling" or "sp," and then requiring students to make the appropriate corrections to improve their scores. Another approach is to establish a "gateway" criterion for all graded written assignments. For example, the teacher specifies that no more than five grammatical, spelling, and punctuation errors will be accepted; if a paper contains more than the specified number, the teacher stops reading and scoring the paper and returns it to the student. The student then corrects the technical errors and resubmits

the paper, possibly for a lower overall score. These methods can be incorporated into any scoring rubric that a nursing faculty member develops for written assignments, as previously discussed.

Suggestions for Assessing and Grading Written Assignments

The suggestions that follow for assessing papers and other written assignments do not apply to every written assignment used in a course, as these are general recommendations to guide teachers in this process.

1. *Relate the assignments to the learning outcomes of the course.* Papers and other written assignments should be planned to meet particular learning outcomes. All too often students complete papers that may have a questionable relationship to course goals.
2. *Consider the number of written assignments to be completed by students, including drafts of papers.* How many care plans; concept maps; subjective, objective, assessment, and plan (SOAP) notes; one-page papers; and so forth are needed to meet the goals of the course? Students should not complete repetitive assignments unless they are essential to meeting course outcomes or objectives, clinical competencies, or personal learning needs.
3. *Avoid assignments that require only summarizing the literature and substance of class and online discussions unless this is the intended purpose of the assignment.* Otherwise students merely report on their readings, often without thinking about the content and how it relates to varied clinical situations. If a review of the literature is the intended outcome, the assignment should direct students to read these articles critically and synthesize them, not merely report on each article.
4. *Include clear directions about the purpose and format of the written assignment.* The goals of the written assignment—why students are writing the paper and how it relates to the course outcomes—should be identified clearly, and generally the more detailed the directions, the better, for both students and the teacher grading the papers. If there is a particular format to be followed, the teacher should review this with students and provide a written or electronic copy for their use in preparing the paper. Students need the criteria for grading and the scoring rubric before they begin the assignment, so it is clear how the paper will be assessed.
5. *Specify the number of drafts to be submitted, each with required due dates, and provide prompt feedback on the quality of the content and writing, including specific suggestions about revisions.* These drafts are a significant component of formal papers because the intent is to improve thinking and writing

through them. Drafts in most instances are used as a means of providing feedback to students and should not be graded.

6. *Develop specific criteria for assessment and review these with the students prior to their beginning the assignment.* The criteria should relate to the quality of the content; organization of content; process of developing ideas and arguments; and elements of writing style such as clarity of expression, sentence structure, punctuation, grammar, spelling, length of the paper, and accuracy and format of the references. Table 9.2 offers a checklist that teachers can use

TABLE 9.2 Checklist for Writing Structure and Style

- ✓ Content organized clearly
- ✓ Each paragraph focuses on one topic and presents details about it
- ✓ Clear sequence of ideas developed within and between paragraphs
- ✓ Clear transitions between paragraphs
- ✓ First sentence of paragraph introduces subject and provides transition from preceding paragraph
- ✓ Paragraphs are of appropriate length
- ✓ Sentences clearly written and convey intended meaning
- ✓ Sentences are of appropriate length
- ✓ Clear transitions between sentences within paragraphs
- ✓ Words express intended meaning and are used correctly
- ✓ Subjects and verbs agree in each sentence
- ✓ Clear antecedents for pronouns
- ✓ No misplaced modifiers
- ✓ Excessive and unnecessary words omitted
- ✓ Stereotypes, impersonal writing, jargon, and abbreviated terms avoided
- ✓ Active voice used
- ✓ Grammar: Correct?
- ✓ Punctuation: Correct?
- ✓ Capitalization: Correct?
- ✓ Spelling: Correct?
- ✓ Writing keeps reader's interest
- ✓ References used appropriately in paper
- ✓ References current
- ✓ No errors in references
- ✓ Correct use of APA or other style for references

Source: Adapted from Oermann, M. H., & Hays, J. (2019). *Writing for publication in nursing* (4th ed.). New York, NY: Springer Publishing Company. Copyright 2019 by Springer Publishing Company. Adapted with permission.

in assessing writing structure and style. Other criteria would be specific to the outcomes to be met through the assignment. If a scoring rubric is used, it should be shared and discussed with the students before they begin the paper.

7. *For papers dealing with analysis of issues, focus the assessment and criteria on the rationale developed for the position taken rather than the actual position.* This type of assignment is particularly appropriate as a group activity in which students critique each other's work.
8. *Read all papers and written assignments anonymously.* The rationale for this is the same as with essay testing—the teacher needs to remove potential bias from the assessment process. Reading papers anonymously helps avoid the chance of a carryover effect in which the teacher develops an impression of the quality of a student's work, for example, from prior papers, tests, or clinical practice, and is then influenced by that impression when grading other assignments. By grading papers anonymously, the teacher also avoids a halo effect.
9. *Skim a random sample of papers to gain an overview of how the students approached the topic of the paper, developed their ideas, and addressed other aspects of the paper that would be graded.* In some cases, the assessment criteria and scoring rubric might be modified, for example, if no students included a particular content area that was reflected in the grading criteria.
10. *Read papers in random order.* Papers read first in the group may be scored higher than those read at the end. To avoid any bias resulting from the order of the papers, it is best to read papers in a random order instead of always organizing papers in the same way (e.g., alphabetical) before reading them. The teacher also should take frequent breaks from grading papers to keep focused on the criteria for evaluation and avoid fatigue, which could influence scoring papers near the end.
11. *Read each paper twice before scoring.* In the first reading, the teacher can note omissions of and errors in content, problems with organization and development of ideas, issues with the process used for developing the paper, and writing-style concerns. For papers submitted electronically, the teacher can insert comments and suggestions in the paper using the Track Changes or Comments tools available in Microsoft Word, or by using different-colored highlighting, making it easy to identify the remarks. For hard copies of papers, comments can be recorded on sticky notes or in pencil in case they need to be modified once the paper is read in its entirety.
12. *If unsure about the assessment of a paper, have a colleague also read and evaluate the paper.* The second reader should review the paper anonymously,

without knowledge of the grade given by the original teacher, and without information about the reason for the additional review. Scores can be averaged, or the teacher might decide to read the paper again depending on the situation. An additional reader also might be used if the grade on the paper will determine whether the student passes the course and progresses in the program. In decisions such as these, it is helpful to obtain a “second opinion” about the quality of the paper.

13. *Consider incorporating student self-critique, peer critique, and group writing exercises within the sequence of writing assignments.* These experiences help students improve their ability to assess their own writing: They can “step back” and reflect on their papers, identify where their ideas may not be communicated clearly, and decide on revisions. Students should be encouraged to ask peers to review and critique their work, similar to asking colleagues to review manuscripts and reports. Group writing activities prepare students for working collaboratively to produce a product, which is similar to nursing practice in real clinical settings and to writing a manuscript as a group.
14. *Prepare students for written assignments by incorporating learning activities in the course, completed in and out of class.* These activities provide practice in organizing and expressing ideas in writing.

■ Summary

Through formal papers, students develop an understanding of the content they are writing about and improve their ability to communicate their ideas in writing. With this type of written assignment, students can analyze and integrate the literature and report on their findings, analyze theories and how they apply to nursing practice, improve their thinking skills, and learn how to write more effectively. To improve their writing abilities, though, students need to complete drafts and rewrites on which they get prompt feedback from the teacher on both content and writing.

There are many types of papers and written assignments that students can complete individually or in small groups in a nursing course. Written assignments should be assessed using predetermined criteria that address quality of content, organization of ideas, the process of arriving at decisions and developing arguments, and writing style. General criteria for evaluating papers, an example of a scoring rubric, and suggestions for assessing and grading written assignments were provided in the chapter.

■ References

- American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.
- Brookhart, S. M., & Nitko, A. J. (2019). *Educational assessment of students* (8th ed.). New York, NY: Pearson Education.
- Bussard, M. E. (2015). Clinical judgment in reflective journals of prelicensure nursing students. *Journal of Nursing Education*, 54, 36–40. doi:10.3928/01484834-20141224-05
- Dahl, H., & Eriksen, K. Å. (2016). Students' and teachers' experiences of participating in the reflection process "THiNK." *Nurse Education Today*, 36, 401–406. doi:10.1016/j.nedt.2015.10.011
- Fulbright, S. (2018, October 18). Using rubrics as a defense against grade appeals. *Faculty Focus*. Retrieved from <https://www.facultyfocus.com/articles/course-design-ideas/rubrics-as-a-defense-against-grade-appeals>
- Halim, A. S., Finkenstaedt-Quinn, S. A., Olsen, L. J., Gere, A. R., & Shultz, G. V. (2018). Identifying and remediating student misconceptions in introductory biology via writing-to-learn assignments and peer review. *CBE Life Sciences Education*, 17(2), ar28. doi:10.1187/cbe.17-10-0212
- Lasater, K. (2011). Clinical judgment: The last frontier for evaluation. *Nurse Education in Practice*, 11, 86–92. doi:10.1016/j.nepr.2010.11.013
- Luthy, K. E., Peterson, N. E., Lassetter, J. H., & Callister, L. C. (2009). Successfully incorporating writing across the curriculum with advanced writing in nursing. *Journal of Nursing Education*, 48, 54–59.
- Minnich, M., Kirkpatrick, A. J., Goodman, J. T., Whittaker, A., Stanton Chapple, H., Schoening, A. M., & Khanna, M. M. (2018). Writing across the curriculum: Reliability testing of a standardized rubric. *Journal of Nursing Education*, 57, 366–370. doi:10.3928/01484834-20180522-08
- Oermann, M. H. (2013). Enhancing writing in online education. In K. H. Frith & D. Clark (Eds.), *Distance education in nursing* (3rd ed., pp. 145–162). New York, NY: Springer Publishing Company.
- Oermann, M. H., & Hays, J. (2019). *Writing for publication in nursing* (4th ed.). New York, NY: Springer Publishing Company.
- Oermann, M. H., Leonardelli, A. K., Turner, K. M., Hawks, S. J., Derouin, A. L., & Hueckel, R. M. (2015). Systematic review of educational programs and strategies for developing students' and nurses' writing skills. *Journal of Nursing Education*, 54, 28–34. doi:10.3928/01484834-20141224-01
- Oermann, M. H., Shellenbarger, T., & Gaberson, K. B. (2018). *Clinical teaching strategies in nursing* (5th ed.). New York, NY: Springer Publishing Company.
- The Joint Commission. (2019). 2019 Hospital National Patient Safety Goals. https://www.jointcommission.org/assets/1/6/2019_HAP_NPSGs_final2.pdf
- Writing Across the Curriculum Clearinghouse. (2019a). What is writing in the disciplines? Retrieved from <https://wac.colostate.edu/resources/wac/intro/wid>
- Writing Across the Curriculum Clearinghouse. (2019b). What is writing to learn? Retrieved from <https://wac.colostate.edu/resources/wac/intro/wtl>



TEST CONSTRUCTION AND ANALYSIS

ASSEMBLING, ADMINISTERING, AND SCORING TESTS

In addition to the preparation of a test blueprint and the skillful construction of test items that correspond to it, the final appearance of the test and the way in which it is administered can affect the validity of the test results. A haphazard arrangement of test items, directions that are confusing, and typographical and other errors on the test may contribute to measurement error. By following certain design rules, teachers can avoid such errors when assembling a test. Administering a test usually is the simplest phase of the testing process. There are some common problems associated with test administration, however, that also may affect the reliability of the resulting test scores and consequently the validity of inferences made about those scores. Careful planning can help the teacher avoid or minimize such difficulties. After administering a test, the teacher's responsibility is to score it or arrange to have it scored. This chapter discusses the processes of assembling the test, administering it to students, and obtaining test scores.

■ Test Design Rules

Allow Enough Time

As discussed in Chapter 3, Planning for Testing, preparing a high-quality test requires time for the design phases as well as for the item-writing phase. Assembling the test is not simply a clerical or technical task; the teacher should make all decisions about the arrangement of test elements and the final appearance of the test even if someone else types or reproduces the test. The teacher must allow enough time for this phase to avoid errors that could affect the students' test scores.

Arrange Test Items in a Logical Sequence

Various methods for arranging items on the test have been recommended, including by order of difficulty and according to the sequence in which the content was taught. However, if the test contains items using two or more formats, the teacher should

first group items of the same format together. Because each item format requires different tasks of the student, this type of arrangement makes it easier for students to maintain the mental set required to answer each type of item, and prevents errors caused by frequent changing of tasks. Keeping items of the same format together also requires fewer sets of directions and facilitates scoring if a scannable answer sheet is not used (Miller, Linn, & Gronlund, 2013). Miller et al. recommended arranging sections of item types in the following order, from simplest to most complex:

1. True–false items
2. Matching exercises
3. Short-answer or completion items
4. Multiple-choice items
5. Context-dependent or interpretive exercises
6. Restricted-response essay items
7. Extended-response essay items (2013, p. 334)

Constructing a test with all of the item types is not recommended, even for a test with a large sample of items administered to a highly skilled group of learners. The longer the test, the more item formats can be included, but complex formats require more reading and processing time for the student, so they should be combined with only one or two other types.

Next, within each item format, items may be arranged according to the order in which the content was taught, which may assist students in recalling information more easily. Finally, combining the item format and content-sequence grouping, teachers should arrange items in order of increasing difficulty. Even well-prepared students are likely to be somewhat anxious at the beginning of a test, and encountering difficult items may increase their anxiety and interfere with their optimum performance. Beginning with easier items may build the students' confidence and allow them to answer these items quickly and reserve more time for difficult items. By having confidence in their ability to answer the beginning items correctly, students may have less anxiety about the remainder of the test (Gronlund, 2006; Miller et al., 2013).

Some nursing faculty members prefer to use test-authoring software to make the process of item writing and test assembly more efficient. However, these applications may or may not allow the teacher to arrange the test items in the manner just described. Many test-development program vendors promote their software's ability to randomize the arrangement of items on a test as a way to produce different versions of the same test to prevent cheating. However, there is no evidence that scrambling the order of items on a test produces a psychometrically equivalent measurement instrument. To facilitate accurate measurement of students' knowledge and skills, all students should respond to the same items arranged in the same order. A later section of this chapter describes more effective ways of preventing cheating on tests.

Write Directions

The teacher cannot assume that the students know the basis on which they are to select or provide answers or how and where to record their answers to test items. Depending on the level of students and their familiarity with the type of items and assessment procedures, it is not reasonable to expect that the assessment will be self-explanatory. This is especially true with students who are non-native English speakers or for those whose primary and secondary education occurred in countries where objectively scored item formats are less common.

The test should begin with a set of clear general directions. These general directions should include instructions on

- How and where to record responses
- What type of writing implement to use
- Whether or not students may write on the test booklet
- The amount of time allowed
- The number of pages and items on the exam
- The types and point values of items
- Whether students may ask questions during the test
- What to do after finishing the exam (Brookhart & Nitko, 2019; Gronlund, 2006; Miller et al., 2013)

Students may need to know some of these instructions while they are preparing for the test, such as whether their answers to items requiring them to supply the names of medications must be spelled accurately to be scored as correct.

Each section consisting of a particular item format should begin with a specific set of instructions. For multiple-choice items, the student needs to know whether to select the *correct* or *best* response. Directions for completion and essay items should state whether spelling, grammar, punctuation, and organization will be considered in scoring, and the length of the desired response. For computation items, directions should specify the degree of precision required, the unit of measure, whether to show the calculation work, and what method of computation to use if there is more than one option (Miller et al., 2013). Matching exercise directions should clearly specify the basis on which the match is to be made (Gronlund, 2006). Here is an example: “For each definition in Column A, select the proper term in Column B. Use each letter in Column B only once or not at all.”

Use a Cover Page

The general test directions may be printed on a cover page (Exhibit 10.1). A cover page also serves to keep the test items hidden from view during the distribution of the exam so that the first students to receive the test will not have more time to

complete it than students who receive their copies later. If the directions on the cover page indicate the number of pages and items, the students can quickly check their test booklets for completeness and correct sequence of pages. The teacher can then replace defective test booklets before students begin answering items.

When a separate answer sheet is used, the cover page may be numbered to help maintain test security; students are directed to record this number in a particular place on the answer sheet. With this system, the teacher can track any missing test booklets after the testing is done. In addition, if the teacher asks students to record responses to short-answer or essay items directly on the test booklet, those answers can be scored anonymously; the score from the answer sheet then can be added to the score from the supply-type items for a total test score that is associated with each student's name.

EXHIBIT 10.1

EXAMPLE OF A COVER PAGE WITH GENERAL DIRECTIONS

Exam Number _____

BEHAVIORAL HEALTH NURSING FINAL EXAM

Directions

1. This test comprises 12 pages. Please check your test booklet to make sure you have the correct number of pages in the proper sequence.
2. Parts I and II contain 86 multiple-choice and matching items. You may write on the test booklet but **you must record your answers to these items on your answer sheet**. This part of the test will be machine scored; read carefully and follow these instructions:
 - a. Use a #2 pencil.
 - b. Notice that the items on the answer sheet are numbered **DOWN** the page in each column.
 - c. Choose the **ONE BEST** response to each item. Items with multiple answer marks will be counted as incorrect. Fill in the circle completely; if you change your answer, erase your first answer thoroughly.
 - d. Print your name (last name, first name) in the blocks provided, then completely fill in the corresponding circle in each column. If you wish to have your score posted, fill in an identification number of up to nine digits (**DO NOT** use your Social Security number) and fill in the corresponding circle in each column.
 - e. Above your name, write your test booklet number.
3. Part III consists of two essay items. Directions for this section are found on page 12. Write your answers to these items on the lined paper provided. You may use pen or pencil. On each page of your answers, write your **TEST BOOKLET NUMBER**. **DO NOT** write your name on these pages.
4. If you have a question during the test, do not leave your seat—raise your hand and a proctor will come to you.
5. You have until 11:00 a.m. to complete this test.

Avoid Crowding

Test items are difficult to read when they are crowded together on the page; non-native English speakers and students with learning disabilities may find crowding particularly trying. Techniques that allow students to read efficiently and to prevent errors in recording their responses include leaving sufficient white space within and between items and indenting certain elements. Teachers should allow enough blank space between and around items so that each item is distinct from the others. If not, the students might inadvertently read a line from a preceding or following item and think it belongs to the item they are answering. Arranging test items on the page in a two-column format also may cause students to lose their place and skip one or more items. Tightly packing words on a page may minimize the amount of paper used for testing, but facilitating maximum student performance on a test is worth a small additional expense for a few more sheets of paper (Miller et al., 2013).

Optimum spacing varies for each item format. The response options for a multiple-choice item should not be printed in tandem fashion, as the following example illustrates:

1. Which method of anesthesia involves injection of an agent into a nerve bundle that supplies the operative site? a. General; b. Local; c. Regional; d. Spinal; e. Topical

The options are much easier to read if listed in a single column below the stem (Miller et al., 2013), as in this example:

1. Which method of anesthesia involves injection of an agent into a nerve bundle that supplies the operative site?
 - a. General
 - b. Local
 - c. Regional
 - d. Spinal
 - e. Topical

Notice in this example that the second line of the stem is indented to the same position as the first line and that the responses are slightly indented. This spacing makes the item number and its content easier to read.

Keep Related Material Together

The stem of a multiple-choice item and all related responses should appear on the same page. Both columns of a matching exercise should also be printed side by side and on one page, including the related directions; using short lists of premises and responses makes this arrangement easier. With context-dependent and interpretive

exercises, the introductory material and all related items should be contained on the same page, if possible. This facilitates reading the material and related questions (Gronlund, 2006; Miller et al., 2013).

Facilitate Scoring

If the test will be scored by hand, the layout of the test or the answer sheet should facilitate easy scoring. A separate answer sheet can be constructed to permit rapid scoring by comparing student responses to an answer key. If the students record their answers directly on the test booklet, the test items should be arranged with scoring in mind. For example, a series of true–false items should be organized with columns of Ts and Fs, preferably at the left margin (Gronlund, 2006; Miller et al., 2013) so that students need to only circle their responses, as in the following example:

- | | |
|-------|--|
| T F | 1. A stethoscope is required to perform auscultation. |
| T F | 2. Physical exam techniques should be performed in the order of least to most intrusive. |
| T F | 3. When using percussion, it is easier to detect a change from dullness to resonance. |

Circling a letter rather than writing or printing it will prevent misinterpretation of the students' handwriting. With completion items, printing blank spaces for the answers in tandem, as in the following example, makes scoring difficult:

1. List three responsibilities of the circulating nurse during induction of general anesthesia.

Instead, the blanks should be arranged in a column along one side of the page, preferably on the left, as in this example:

- | | |
|--|---|
| <ol style="list-style-type: none"> 1. _____ 2. _____ 3. _____ | <ol style="list-style-type: none"> 1–3. List three responsibilities of the circulating nurse during induction of general anesthesia. |
|--|---|

Arrange the Correct Answers in a Random Pattern

Many teachers have a tendency to favor certain response positions for the correct or keyed answer to objective test items, for example, to assign the correct response to

the A or D position of a multiple-choice item. Some teachers arrange test items so that the correct answers form a pattern that makes scoring easy (e.g., T-F-T-F or A-B-C-D). Students who detect a pattern of correct answers (e.g., the correct answer never appears in the same position two or more consecutive times) may use this information to obtain higher test scores than their knowledge would warrant (Gronlund, 2006).

Many item analysis software programs calculate the number of times the keyed response occurs in each position. While reviewing these reports, teachers may determine whether the correct answer positions occur in approximately equal numbers, keeping in mind that multiple-choice, true-false, and matching items may have differing numbers of response options. Although these reports would not be available until after the test is administered and scored, they could alert teachers to use a different technique to more evenly distribute the correct answer position if the test (in its entirety or with minor item revisions) is used again. The teacher also can tally the number of Ts and Fs, or As, Bs, Cs, and Ds, on the answer key by hand. For true-false items, if either true or false statements are found to predominate, some items may be rewritten to make the distribution more equal (although it is recommended by some experts to include more false than true items).

Gronlund (2006) recommended that the position of the correct response in multiple-choice items be randomly assigned. One method for obtaining a random order is to place all responses to multiple-choice items and all premises and responses in a matching exercise in alphabetical order by the first letter in each, as described in the following section.

Arrange Options in Logical or Numerical Order

The response alternatives for multiple-choice items and the premises and responses of a matching exercise should be arranged according to a logical or meaningful order, such as alphabetical or chronological order, or in order of size or degree. This type of arrangement reduces reading time and helps students who know the correct answer to search through the options to find it. This strategy also tends to randomly distribute the correct answer position as discussed earlier, especially on lengthy tests. When the options are numbers, they should always be in numerical order, preferably ascending (Gronlund, 2006). This principle can be seen in the example shown in Exhibit 10.2.

Number the Items Consecutively Throughout the Test

Although test items should be grouped according to format, they should be numbered consecutively throughout the test. That is, the teacher should not start each new item format section with item number 1 but continue numbering items in a continuous

EXHIBIT 10.2

ARRANGEMENT OF OPTIONS: NOT ORDERED VERSUS ORDERED NUMERICALLY

OPTIONS NOT ORDERED	OPTIONS IN NUMERICAL ORDER
Your patient is ordered guaifenesin 300 mg four times daily. It comes 200 mg/5 mL. How many milliliters should you give per dose?	Your patient is ordered guaifenesin 300 mg four times daily. It comes 200 mg/5 mL. How many milliliters should you give per dose?
a. 5.0 mL b. 2.5 mL c. 10 mL d. 7.5 mL ¹	a. 2.5 mL b. 5.0 mL c. 7.5 mL ¹ d. 10 mL

Note: ¹ = correct answer.

sequence. This numbering system helps students to find items they may have skipped and to avoid making errors when recording their answers, especially when using a separate answer sheet.

Proofread

The goal throughout the preparation and use of assessments is to obtain valid evidence that students have met learning goals. Although validity is a major focus of the planning for a test (e.g., through use of a test blueprint), careful assembly and administration of the test will ensure that it will function as intended (Miller et al., 2013).

The test items and directions should be free of spelling, punctuation, grammatical, and typing errors. Such defects are a source of measurement error and can cause confusion and distraction, particularly among students who are anxious (Brookhart & Nitko, 2019). Typographical and similar errors are a problem for any student but more so for non-native English speakers or those who have learning disabilities. Often the test designer does not recognize his or her own errors; another teacher who knows the content may be asked to proofread a copy of the test before it is duplicated. The spell-check or grammar-check features of a word processing program may not recognize punctuation errors or words that are spelled correctly but used in the wrong context, and they may not always detect structural errors such as giving two test items the same number or two responses the same letter.

Prepare an Answer Key

Whether the test will be machine-scored or hand-scored, the teacher should prepare and verify an answer key in advance to facilitate efficient scoring and to provide a final check on the accuracy of the test items. Scannable answer sheets also can be

used for hand-scoring; an answer key can be produced by punching holes to indicate the correct answers. The teacher also should prepare ideal responses to essay items, identify intended responses to completion items, and prepare scoring rubrics if the analytical scoring method is used.

■ Reproducing the Test

Ensure Legibility

Legibility is an important consideration when printing and duplicating the test; poor-quality copies may interfere with optimum student performance. A font that includes only uppercase letters is difficult to read; upper- and lowercase lettering is recommended. The master or original copy should be letter quality, produced with a high-quality printer so that it can be clearly reproduced. For best results, the test should be photocopied or printed on a machine that has sufficient toner to produce crisp, dark print without any stray lines or artifacts.

Print on One Side of the Page

The test should be reproduced on only one side of each sheet of paper. Printing on both sides of each page could cause students to skip items unintentionally or make errors when recording their scores on a separate answer sheet. It also creates distractions from excessive page-turning during the test. If the test is to be hand-scored and students record their answers on the test rather than on a separate answer sheet, printing only on one side makes it easier to score.

Reproduce Enough Copies

The teacher should duplicate more test copies than the number of students to allow for extra copies for proctors or to replace defective copies that may have been inadvertently distributed to students. Displaying test items on a screen using a projector, or writing them on the chalkboard or interactive whiteboard may save costs or the teacher's preparation time, but these procedures may cause problems for students with learning or visual disabilities. When students do not have their own copies of a test for whatever reason, they cannot control the pace at which they answer items or return to a previous item. Dictating test items is not recommended except when the objective is to test knowledge of correct spelling; in addition to creating problems for students with hearing impairments, this method wastes time that students could otherwise spend in thinking about and responding to the items. In addition, there would be no record of how the items were worded, which could present a problem if a student later questions how an answer was scored.

Maintain Test Security

Teachers have an important responsibility to maintain the security of tests by protecting them from unauthorized access. Carelessness on the part of the teacher can enable dishonest students to gain access to test materials and use them to obtain higher scores than they deserve. This contributes to measurement error, and it is unfair to honest students who are well-prepared for the test. It is up to the teacher to make arrangements to secure the test while it is being prepared, reproduced, stored, administered, and scored.

Test materials should be stored in locked areas accessible only to authorized personnel. Computer files that contain test items should be protected with passwords, encryption, or similar security devices. Only regular employees should handle test materials; student employees should not be asked to type, print, or reproduce tests. While test items are being typed, they should be protected from the view of others. Printed drafts of tests should be destroyed by shredding pages rather than discarding them in trash or recycling receptacles.

As previously mentioned, one suggestion for preventing cheating during test administration to large groups is to prepare alternative forms of the test. This can be done by presenting the same questions but in a different order on each form. However, the psychometric properties of alternative forms produced in these ways might be sufficiently different as to result in different scores, especially when the positions of items with unequal difficulty are switched. For calculation items, the teacher can modify values within the same question on different forms; in that way the responses will not be identical. Similarly, the order of responses to multiple-choice and matching items might be scrambled to produce an alternative form of the test. If there is little or no evidence for the true equivalence of these alternative forms, it is best not to use this approach. Other ways to prevent cheating are discussed in the next section of this chapter.

■ Test Administration

Environmental Conditions

The environmental conditions of test administration can be a source of measurement error if they interfere with the students' performance. If possible, the teacher should select a room that limits potential distractions during the test. For example, if windows must be open for ventilation during warm weather, the students may be distracted by lawn mowing or construction noise; requesting a room for testing on another side of the building may prevent the problem. Placing a sign such as "Testing—Quiet Please" on the door of the classroom may reduce noise in the hallway (Miller et al., 2013).

For online courses, it is critical to determine prior to the test administration that students have the computer capabilities and Internet access to take the exam for the time period allotted. Students with dial-up modems may experience “timing out,” which means they are being disconnected from the Internet by their Internet service providers after a set period of time or what appears to be inactivity on the part of the user. When that occurs, the students cannot transmit their completed exams, and course management systems may not permit them to access another copy. A more extensive discussion of effective approaches to online testing can be found in Chapter 11, *Testing and Evaluation in Online Courses and Programs*.

Distributing the Test Materials

Careful organization allows the teacher to distribute test materials and give instructions to the students efficiently. With large groups of students, several proctors may be needed to assist with this process. If a separate answer sheet is used, it usually can be distributed first, followed by the test booklets. During distribution of the test booklets, the teacher should instruct students not to turn over the test booklet and begin the test until told to do so. At this point, the students should check their test booklets for completeness, and the proctors should replace defective booklets. The teacher then should read the general directions aloud while the students read along. Hearing the directions may help non-native English speakers, students with learning disabilities, and students whose anxiety may interfere with their comprehension of the written instructions. Once the teacher answers any general questions about the test procedures, the students can begin the test.

However, do not take any more time than necessary before allowing students to begin the test. Extended remarks and instructions may interfere with students’ mental set for the test, increase students’ anxiety, and possibly create hostility toward the teacher (Miller et al., 2013).

Answering Questions During the Test

Some students may find it necessary to ask questions of the teacher during a test, but responding to these questions is always somewhat disturbing to other students. Also, by responding to student questions during a test, a proctor may inadvertently give hints to the correct answer, which would put that student at an advantage while not making the same information available to other students (Miller et al., 2013). However, it is not appropriate to refuse to allow questions during a test. One of the teacher’s responsibilities in administering a test professionally is to provide “reasonable opportunities for individuals to ask questions about the assessment procedures or directions prior to and at appropriate times during administration” (National Council on Measurement in Education, 1995, Section 4.9). If a student asks

a question that the proctor cannot answer, the student may be instructed to record the question on a separate piece of paper identified with the student's name; questions can be collected with the other test materials. Then if a student identifies a flaw in a test item, the teacher can take the necessary action after the test is completed rather than interrupt the test to announce corrections. Chapter 12, Test and Item Analysis: Interpreting Test Results, includes a discussion of how to adjust test scores if an item is found to be fatally flawed.

While answering student questions during the test, distraction can be kept to a minimum by telling students to raise their hands if they have questions rather than leaving their seats to approach the teacher; a proctor then goes to the student's seat. Proctors should answer questions as quietly and briefly as possible. In answering questions, proctors certainly should address errors in the test copy and ambiguity in directions but should avoid giving clues to the correct answers. When writing items, teachers should work to eliminate cultural bias and terms that would be unfamiliar to non-native English speakers. This is discussed further in Chapter 16, Social, Ethical, and Legal Issues.

Preventing Cheating

Cheating is widely believed to be common on college campuses in the United States. Brookhart and Nitko (2019) suggested that when teachers know their students, interact with them about their learning, and give meaningful assignments, they create an environment in which cheating is less likely to occur.

Cheating is defined as any activity whose purpose is to gain a higher score on a test or other academic assignment than a student is likely to earn on the basis of achievement. Traditional forms of cheating on a test include but are not limited to the following:

- Acquiring test materials in advance of the test or sharing materials with others
- Arranging for a substitute to take a test
- Preparing and using unauthorized notes or other resources during the test
- Exchanging information with others or copying answers from another student during the test
- Copying test items or retaining test materials to share with others who may take the test later

In addition to the low-technology forms of cheating on a test such as writing on body parts, clothing (e.g., the underside of the bill of a cap, the inside of a sleeve or waistband), or belongings (e.g., backpack, jewelry, facial tissue) and copying answers from others, technological advances have created many new, more sophisticated

methods. For example, students with smartphones and -watches can transmit information to other students or solicit help from them via messaging, email, and camera. Cell phones and other “smart” devices are easily concealed by students under desks or in baggy clothing. The widespread use of Bluetooth technology makes this practice even easier. Students with MP3 players or similar devices can listen to pre-recorded content related to the domain being tested—a sort of auditory cheat sheet. Teachers who allow students to use handheld devices during a test to access tools helpful in solving problems (e.g., calculators for solving medication dosage calculation problems) must be especially vigilant. The faculty member should fully understand the functions of such devices to curb such practices as preprogramming and use of multiple screens that can be minimized (Hulsart & McCarthy, 2009).

With adequate test security and good proctoring during the test, the teacher usually can prevent these opportunities for cheating. Students who do act honestly resent those who cheat, especially if dishonest students are rewarded with high test scores. Honest students also resent faculty members who do not recognize and deal effectively with cheating.

Because of the widespread and growing use of technological aids to cheating, teachers should consider instituting standard procedures to be followed during all tests, especially if testing large groups of students. Included in these procedures may be conditions such as the following:

- No personal belongings may be brought into the testing room other than a writing implement. Backpacks; books; papers; cell phones, pagers, and other handheld devices; purses; briefcases; tissues; candy or cough drops; beverage bottles or cups; “lucky charms”; and so forth, must be left outside the classroom.
- Outerwear such as coats, jackets, and caps with a bill or brim may not be worn.
- Sunglasses or visors may not be worn.
- Earplugs or earbuds may not be worn. If students wish to use earplugs to block environmental noise during tests, they should inform the teacher in advance, and the teacher may supply inexpensive, disposable ones.
- The teacher may provide a supply of scratch paper to be used during the test and submitted with other test materials before students leave the testing room.
- The teacher may provide a supply of tissues and extra writing implements to be used if needed during the test.
- Bathroom breaks during the test may be prohibited or limited, depending on the testing time allowed. Students may need to be accompanied to restrooms by proctors, who may search restrooms for hidden devices and print resources before students are permitted to use them.

- Students will occupy every other seat in a row, directly behind students in the row in front of them.
- Students must keep test materials on the desk or table in full view of the proctors and not spread out over a large area. If a student must leave the testing room for any reason, all test materials should be turned facedown during the student's absence.
- If the use of calculators is permitted during exams, the faculty may purchase the necessary quantity of an inexpensive model with limited functionality to be distributed and collected with the test materials.
- Students may not leave their seats without permission until they have completed the test and are submitting their test materials.

Although some of these measures may appear extreme, many of them are variations of the test conditions under which graduates of the nursing education program will take licensure or certification examinations. Students may benefit from becoming accustomed to taking tests under these conditions. Teachers should decide which, if any, of these suggestions are appropriate for use in their particular circumstances.

Although a number of methods for preventing cheating during a test have been proposed, the single most effective method is careful proctoring. There should be enough proctors to supervise students adequately during exams; for most groups of students, at least two proctors are suggested so that one is available to leave the room with a student in case of emergency without leaving the remaining students unsupervised. When proctoring a test, it is important to be serious about the task and devote full attention to it rather than grading papers, checking email and other messages, or reading. If more than one proctor is available, they should locate themselves at different places in the room to observe students from different vantage points. Proctors should avoid walking around the room unless in response to a student's raised hand; such walking can be distracting, especially to students with test anxiety.

A particularly troubling situation for teachers is how to deal with a student's behavior that suggests cheating during a test. Prior to administering the test, the teacher must know the policies of the nursing education program and college or university regarding cheating on an examination or another assessment. If a teacher is certain that a student is cheating, the teacher should quietly collect the test and answer sheet and ask the student to leave the room. However, if it is possible that the teacher's interpretation of the behavior is incorrect, it may be best not to confront the student at that time. In addition to preventing a potentially innocent student from completing the test, confiscating test materials and ordering a student to leave will create a distraction to other students that may affect the accuracy of all the students'

test scores. A better response is to continue to observe the student, making eye contact if possible to make the student aware of the teacher's attention. If the student was attempting to cheat, this approach usually effectively stops the behavior. If the behavior continues, the teacher should attempt to verify this observation with another proctor, and if both agree, the student may be asked to leave the room.

Although many testing experts would argue that the appropriate penalty for cheating on a test is a score of zero for that test, Brookhart and Nitko (2019) referred to this approach as “the deadly zero” (pp. 342–343). Depending on the number of components that contribute to the course grade and the relative weight of each, a test score of zero as a consequence of cheating may result in a failing grade for the course. (See Chapter 17, Grading, for a more comprehensive discussion of grading components.) However, simply deducting a predetermined number of points from the test score suggests that the low score represents the offending student's true level of achievement, which is not the case. Brookhart and Nitko discussed several strategies for computing a course grade when one component was missing (an assignment that was not submitted); in one of the strategies, the teacher assigns the highest possible failing score according to the grading scale in use instead of a zero, which tends to have a less devastating effect on the course grade. Although their recommendations were made in the context of a missing assignment, the same principles might be applied to the question of an appropriate sanction for cheating on a test. Whatever strategy teachers choose as a sanction for cheating on a test, they are using grades to control students' behavior by lowering a score that is meant to indicate achievement “for behavior that is unrelated to achievement” (Brookhart & Nitko, 2019, p. 341). The sanction for cheating on a test should be specified in an academic honesty policy that is consistent with that of the parent institution, and students should be informed of the policy before it is enforced.

If the teacher learns that a copy of a test is circulating in advance of the scheduled date of administration, the teacher should attempt to obtain verifiable evidence that some students have seen it. In this case, the teacher needs to prepare another test or develop an alternative way of assessing student learning. As described in this book, there are many assessment strategies applicable for measuring learning outcomes in nursing.

Online Testing

As more courses and programs are offered through distance education, teachers are faced with how to prevent cheating on an assessment when they cannot directly observe the students. Various approaches can be used, ranging from administering the tests in a proctored computer testing center to high-technology solutions such as remote proctoring. A more extensive discussion of this topic can be found in Chapter 11, Testing and Evaluation in Online Courses and Programs.

Collecting Test Materials

For traditional on-site tests, when students are finished with the test and are preparing to leave the room, the resulting confusion and noise can disturb students who are still working. The teacher should plan for efficient collection of test materials to minimize such distractions and to maintain test security. It is important to be certain that no test materials leave the room with the students. Therefore, teachers should take care to verify that the students turn in their test booklets, answer sheets, scratch paper, and any other test materials. With a large group of students, one proctor may be assigned the task of collecting test materials from each student; this proctor should check the test booklet and answer sheet to ensure that the directions for marking answers were followed, that the student's name (or number) is recorded as directed, and that the student has not omitted any items. Any such errors can then be corrected before the student leaves the room, and test security will not be compromised.

If students are still working near the end of the allotted testing time, the remaining amount of time should be announced, and they should be encouraged to finish as quickly as possible. When the time is up, all students must stop, and the teacher or proctor must collect the rest of the tests. Students who have not finished the test at that point cannot have additional time unless they have been granted that specific accommodation for qualified learning disabilities. This determination should be made in advance of the test and the necessary arrangements made. Extended testing time is not an appropriate remedy for every learning disability, however. It should be provided only when specifically prescribed based on a psychoeducational evaluation of a student's abilities and needs. Chapter 16, Social, Ethical, and Legal Issues, includes additional discussion of accommodations for students with disabilities.

Collaborative Testing

Collaborative testing, an assessment method in which pairs or small groups of students work together during summative assessments, is gaining support from both teachers and students at all educational levels. There are a number of collaborative testing methods, but most involve students taking the same test twice: once individually, and then, after submitting their answer sheets, meeting in small groups to discuss the test items and then retake the test. In most cases, pairs or small groups are randomly assigned at the time of the test. The manner of retesting varies; in some methods, dyads or small groups discuss the test items but submit separate answer sheets, resulting in individual scores. In this procedure, students are not required to answer on the basis of group consensus or vote on the answer to each item, but instead record their own answers after the discussion. Teachers may

record the sum or mean scores of the two individual tests. In other methods, the pairs or groups discuss the test items until they reach consensus on the answers, and one answer sheet is submitted for the pair or group. Each student's total score for the two tests is then determined by some combination of the individual and the group scores, for example, the sum, the mean of the two scores, requiring a passing score on the individual test before receiving additional points from a collaborative test score, or a weighted score such as 2/3 of the individual score and 1/3 of the collaborative score (Burgess & Medina-Smuck, 2018; Rivaz, Momennasab, & Shokrollahi, 2015).

Studies of collaborative testing in chiropractic and nursing education programs have demonstrated better performance in the collaborative testing groups and student preference for collaborative testing. In both research and anecdotal reports, students have consistently reported positive perceptions of collaborative testing, including decreased test anxiety, improved thinking skills, and increased motivation. In a study of the impact of collaborative testing in a graduate nursing program, Phillips, Munn, and George (2019) found it to be an effective strategy for improving teamwork and communication skills, enhancing relationships, and facilitating critical thinking. By encouraging students to participate as active learners, collaborative testing may support positive attitudes about the importance of course content, enhance depth of learning, and improve higher level thinking skills (Meseke, Nafziger, & Meseke, 2010; Sandahl, 2010). However, reported effects of collaborative testing on longer term knowledge retention have not been consistent. Rivaz et al. (2015) found a significant improvement in retention of medical–surgical nursing content among undergraduate students who had taken both individual and collaborative tests compared to students who had taken only the individual test, but other studies found no difference in long-term knowledge retention between collaborative and traditional testing (Sandahl, 2010). Results also may vary according to the cognitive level being measured by the test, with students performing better on collaborative tests with relatively low-level items. Collaborative testing apparently benefits both low- and high-performing students, but low performers have shown significantly higher group test than individual test scores. Students involved with collaborative testing have reported studying no more than they would have normally but demonstrated better overall course performance as compared with students involved in traditional solo testing (Meseke et al., 2010).

Despite the reported benefits of collaborative testing, students in some studies have reported concern that their unprepared classmates may have earned higher exam scores than they deserved. It has been noted that some individuals contribute little to the collaborative efforts while reaping the benefits of the group interaction, a phenomenon known as *social loafing*, *free riding*, or *freeloading*. Although this behavior may disrupt group functioning, it may appear to be advantageous to

low-achieving students who receive input from their peers without reciprocating. However, “parasitic” students who do not participate fully in the discussion may not learn as deeply as those who do, and even though high-achieving students may be annoyed by this behavior, it probably does them little harm because they will benefit from the discussion and group feedback. The freeloading problem is less problematic in smaller sized groups (no more than four students) due to a level of peer pressure that promotes participation (Meseke et al., 2010).

Students may complain about the inclusion of higher cognitive level items on collaborative tests because of difficulty in reaching consensus about the correct answers; students’ individual answers on the retest do not always correlate with the answer recommended by the group after discussion. Students also have reported that lower level items did not enhance their critical thinking skills because the group was able to reach consensus quickly without much discussion (Meseke et al., 2010; Sandahl, 2010).

Collaborative testing typically is used for only some of the tests in a nursing course, most often selected quizzes and unit exams. Although students may benefit in a number of ways from this testing method, they still must develop sufficient skill at taking individual tests to support their success on licensure and certification examinations that they will take on completion of the nursing education program.

■ Scoring

Many teachers say that they “grade” tests, when in fact it would be more accurate to say that they “score” tests. *Scoring* is the process of determining the first direct, unconverted, uninterpreted measure of performance on a test, usually called the *raw*, *obtained*, or *observed score*. The raw score represents the number of correct answers or number of points awarded to separate parts of an assessment (Brookhart & Nitko, 2019). On the other hand, *grading* or *marking* is the process of assigning a symbol to represent the quality of the student’s performance. Symbols can be letters (A, B, C, D, F, which may also include + or –), categories (pass–fail, satisfactory–unsatisfactory), or percentages (100, 99, 98, . . .), among other options.

In most cases, test scores should not be converted to grades for the purpose of later computing a final average grade. Instead, the teacher should record actual test scores and then combine all scores into a composite score that can be converted to a final grade. Recording scores contributes to greater measurement accuracy because information is lost each time scores are converted to symbols. For example, if scores from 70 to 79 all are converted to a grade of C, each score in this range receives the same grade, although scores of 71 and 78 may represent important differences in achievement. If the C grades all are converted to the same numerical grade, for example, C = 2.0, then such distinctions are lost when the teacher computes the final grade for the course. Various grading systems and their uses are discussed in Chapter 17, Grading.

Weighting Items

As a general rule, each objectively scored test item should have equal weight. Most electronic scoring systems assign 1 point to each correct answer unless the teacher specifies a different item weight; this seems reasonable for hand-scored tests as well. It is difficult for teachers to justify that one item is worth 2 points whereas another is worth 1 point; such a weighting system also motivates students to argue for partial credit for some answers.

Differential weighting implies that the teacher believes knowledge of one concept is more important than knowledge of another concept. When this is true, the better approach is to write more items about the important concept; this emphasis would be reflected in the test blueprint, which specifies the number of items for each content area. When a combination of selection-type items and supply-type items is used on a test, a variable number of points can be assigned to short-answer and essay items to reflect the complexity of the required task and the value of the student's response (Miller et al., 2013). It is not necessary to adjust the numerical weight of items to achieve a total of 100 points. Although a test of 100 points allows the teacher to calculate a percentage score quickly, this step is not necessary to make valid interpretations of students' scores.

Correction for Guessing

With standardized tests, the raw score sometimes is adjusted or corrected before it is interpreted. One procedure involves applying a formula intended to eliminate any advantage that a student might have gained by guessing correctly. The correction formula reduces the raw score by some fraction of the number of the student's wrong answers (Brookhart & Nitko, 2019; Miller et al., 2013). The formula can be used only with simple true-false, multiple-choice, and some matching items, and is dependent on the number of alternatives per item. The general formula is

$$\text{Corrected score} = R - \frac{W}{n-1} \quad (10.1)$$

where R is the number of right answers, W is the number of wrong answers, and n is the number of options in each item (Miller et al., 2013). Thus, for two-option items like true-false, the teacher merely subtracts the number of wrong answers from the number of right answers (or raw score); for four-option items, the raw score is reduced by one third of the number of wrong answers. A correction formula is obviously difficult to use for a test that contains several different item formats.

The use of a correction formula usually is appropriate only when students do not have sufficient time to complete all test items and when they have been instructed not to answer any item for which they are uncertain of the answer (Miller et al., 2013). Even under these circumstances, students may differ in their interpretation of "certainty"

and therefore may interpret the advice differently. Some students will guess regardless of the instructions given and the threat of a penalty; the risk-taking or testwise student is likely to be rewarded with a higher score than the risk-avoiding or non-testwise student because of guessing some answers correctly. These personality differences cannot be equalized by instructions not to guess and penalties for guessing.

The use of a correction formula also is based on the assumption that the student who does not know the answer will guess blindly. However, Brookhart and Nitko (2019) suggested that the chance of getting a high score by random guessing is slim, though many students choose correct answers through informed guesses based on some knowledge of the content. Based on these limitations and the fact that most tests in nursing education settings are not speeded, the best approach is to advise all students to answer every item, even if they are uncertain about their answers, and apply no correction for guessing.

■ Summary

The final appearance of a test and the way in which it is administered can affect the validity of the test results. Poor arrangement of test items, confusing or missing directions, typographical errors, and careless administration may contribute to measurement error. Careful planning can help the teacher to avoid or minimize these difficulties.

Rules for good test design include allowing sufficient time, arranging test items in a logical sequence, writing general and item-format directions, using a cover page, spacing test elements to avoid crowding, keeping related material together, arranging the correct answers in a random pattern, numbering items consecutively throughout the test, proofreading the test, and preparing an accurate answer key. In preparing to reproduce the test, the teacher should ensure legibility, print the test on one side of each page, prepare enough copies for all students and proctors, and maintain the security of test materials.

Although administering a test usually is the simplest phase of the testing process, there are some common problems that may affect the reliability of the resulting scores. Teachers should arrange for favorable environmental conditions, distribute the test materials and give directions efficiently, make appropriate plans for proctoring and answering questions during the test, and collect test materials efficiently. Teachers have an important responsibility to prevent cheating before, during, and after a test. Various forms of cheating were discussed, and suggestions were given for preventing cheating on a test, including careful proctoring.

The chapter also included a brief discussion of collaborative testing. Several methods of this testing paradigm were described. Studies have reported satisfaction with collaborative testing, but some expressed concern about unprepared peers

and those who contribute little to group discussion receiving higher test scores than they deserved. In general, collaborative testing appears to benefit both high- and low-achieving students, but probably should not be used for all tests whose scores will contribute to course grades.

After administering a test, the teacher must score it and interpret the results. Scoring is the process of determining the first direct, uninterpreted measure of performance on a test, usually called the *raw score*. The raw score usually represents the number of right answers. Test scores should not be converted to grades for the purpose of later computing a final average grade. Instead, the teacher should record actual test scores and then combine them into a composite score that can be converted to a final grade.

As a general rule, each objectively scored test item should have equal weight. If knowledge of one concept is more important than knowledge of another concept, the teacher should sample the more important domain by writing additional items in that area. Most machine-scoring systems assign 1 point to each correct answer; this seems reasonable for hand-scored tests as well.

A raw score sometimes is adjusted or corrected before it is interpreted. One procedure involves applying a formula intended to eliminate any advantage that a student might have gained by guessing correctly. Correcting for guessing is appropriate only when students have been instructed not to answer any item for which they are uncertain of the answer; students may interpret and follow this advice differently. Therefore, the best approach is to advise all students to answer every item, with no correction for guessing applied.

■ References

- Brookhart, S. M., & Nitko, A. J. (2019). *Educational assessment of students* (8th ed.). New York, NY: Pearson Education.
- Burgess, A., & Medina-Smuck, M. (2018). Collaborative testing using quizzes as a method to improve undergraduate nursing student engagement and interaction. *Nursing Education Perspectives*, 39, 178–179.
- Gronlund, N. E. (2006). *Assessment of student achievement* (8th ed.). Boston, MA: Pearson Education.
- Hulsart, R., & McCarthy, V. (2009). Educators' role in promoting academic integrity. *Academy of Educational Leadership Journal*, 13(4), 49–61.
- Meseke, C. A., Nafziger, R., & Meseke, J. K. (2010). Student attitudes, satisfaction, and learning in a collaborative testing environment. *Journal of Chiropractic Education*, 24, 19–29.
- Miller, M. D., Linn, R. L., & Gronlund, N. E. (2013). *Measurement and assessment in teaching* (11th ed.). Upper Saddle River, NJ: Prentice Hall.
- National Council on Measurement in Education. (1995). *Code of professional responsibilities in educational measurement (CPR)*. Retrieved from <https://www.ncme.org/resources/library/professional-responsibilities>

- Phillips, T. A., Munn, A. C., & George, T. P. (2019). The impact of collaborative testing in graduate nursing education. *Journal of Nursing Education*, 58, 357–359. doi:10.3928/01484834-20190521-07
- Rivaz, M., Momennasab, M., & Shokrollahi, P. (2015). Effect of collaborative testing on learning and retention of course content in nursing students. *Journal of Advances in Medical Education & Professionalism*, 3, 178–182.
- Sandahl, S. S. (2010). Collaborative testing as a learning strategy in nursing education. *Nursing Education Perspectives*, 11, 142–147.

TESTING AND EVALUATION IN ONLINE COURSES AND PROGRAMS

Contemporary nursing students expect educational institutions to provide flexible instructional methods that help them balance their academic, employment, family, and personal commitments (Jones & Wolf, 2010). Online education has rapidly developed as a potential solution to these demands. The growth rate of online student enrollment in all disciplines has far exceeded the growth rate of traditional course student enrollment in U.S. higher education (Seaman, Allen, & Seaman, 2018). Over 6.3 million students enrolled in at least one college-level online course during the fall 2016 academic term, with the proportion of all students taking at least one online course representing 32.0% of all students (Seaman et al., 2018). In nursing, the American Association of Colleges of Nursing (2019) reported that some of the 219 RN-to-master's degree programs and more than 600 RN-to-BSN programs were offered at least partially online. Doctor of nursing practice (DNP) programs, which have experienced a huge growth over the past few years, are frequently offered partially or completely online, and online courses are included in prelicensure and other levels of nursing education programs as well.

For the purposes of this chapter, *online courses* are those in which at least 80% of the course content is delivered online. Face-to-face courses are those in which 0% to 29% of the content is delivered online; this category includes both traditional- and web-facilitated courses. Blended (sometimes called *hybrid*) courses have between 30% and 80% of the course content delivered online. Examples of various course management systems used for online courses include Blackboard, Desire2Learn (Brightspace), Sakai, and Moodle.

Along with the expansion of online delivery of courses and programs comes concern about how to evaluate their quality. Chapter 18, Program Evaluation and Accreditation, provides standards for evaluating distance education programs. The accreditation criteria from each of the accrediting bodies in nursing address online programs in different ways, and these are presented in Chapter 18, Program Evaluation and Accreditation. This chapter discusses recommendations for assessment of

learning in online courses, including testing and appraising course assignments, to determine whether course goals and outcomes have been met. It also suggests ways to assess online courses and programs, and to assess teaching effectiveness in online courses and programs.

■ Assessment of Learning at the Individual Learner Level

Online assessment and evaluation principles do not differ substantially from the approaches used in the traditional classroom environment. As with traditional format courses, assessment of individual achievement in online courses should involve multiple methods such as tests, written assignments, and contributions to online discussions. Technological advances in testing and assessment have made it possible to administer tests on a computer and assess other products of student thinking even in traditional courses (Miller, Linn, & Gronlund, 2013). But courses and programs that are offered only online or in a hybrid format depend heavily or entirely on technological methods to assess the degree to which students have met expected learning targets or outcomes.

Online Testing

The choice to use online testing inevitably raises concerns about academic dishonesty. How can the course instructor be confident that students who are enrolled in the course are the ones who are taking the tests? How can teachers prevent students from consulting unauthorized sources while taking tests or sharing information about tests with students who have not yet taken them? To deter cheating and promote academic integrity, faculty members should incorporate a multifaceted approach to online testing. Educators can employ low- and high-technology solutions to address this problem.

One example of a low-technology solution includes creating an atmosphere of academic integrity in the classroom by including a discussion of academic integrity expectations in the syllabus or student handbook (Conway-Klaassen & Keil, 2010; Hart & Morgan, 2009). When teachers have positive relationships with students, interact with them regularly about their learning, and convey a sense of confidence about students' performance on tests, they create an environment in which cheating is less likely to occur (Brookhart & Nitko, 2019; Miller et al., 2013). Faculty members should develop and communicate clear policies and expectations about cheating on online tests, plagiarism, and other examples of academic dishonesty (Morgan & Hart, 2013). Unfortunately, students do not always view cheating or sharing as academic dishonesty; they often believe it is just collaboration (Wideman, 2011).

Another low-technology option is administering a tightly timed examination. This approach may deter students from looking up answers to test items for fear of running out of time to complete the assessment. Other suggestions to minimize cheating on online examinations include randomizing the test items and response options, displaying one item at a time and not allowing students to review previous items and responses, creating and using different versions of the test for the same group of learners, and developing open-book examinations (Conway-Klaassen & Keil, 2010). However, each of these approaches has disadvantages that teachers of online courses must take into consideration before implementing them.

Randomized Sequence of Test Items and Response Options

As discussed in Chapter 10, Assembling, Administering, and Scoring Tests, the sequence of test items may affect student performance and therefore assessment validity. Many testing experts recommend arranging items of each format in order of difficulty, from easiest to most difficult, to minimize test anxiety and allow students to respond quickly to the easy items and spend the majority of testing time on the more difficult ones. Another recommendation is to sequence test items of each format in the order in which the content was taught, allowing students to use the content sequence as a cognitive map by which they can more easily retrieve stored information. A combination of these approaches—content sequencing with difficulty progression within each content area—may be the ideal design for a test (Brookhart & Nitko, 2019). Many testing experts also recommend varying the position of the correct answer to multiple-choice and matching items in a random way to avoid a pattern that may help testwise but uninformed students achieve higher scores than their knowledge warrants. A simple way to obtain sufficient variation of correct answer position is to arrange the responses in alphabetical or numerical order (Brookhart & Nitko, 2019; Gronlund, 2006). Therefore, scrambling the order of test items and response options on an online test may affect the validity of interpretation of the resulting scores, and there is no known scientific evidence to recommend this practice as a way of preventing cheating on online tests.

Displaying One Item at a Time and Not Allowing Students to Review Previous Items

This tactic is appropriate for the computerized adaptive testing model in which each student's test is assembled interactively as the person is taking the test. Because the answer to one item (correct or incorrect) determines the selection of the next item, there is nothing to be gained by reviewing previous items. However, in teacher-constructed assessments for traditional or online testing, students should

be permitted and encouraged to return to a previous item if they recall information that would prompt them to change their responses. While helping students develop test-taking skills to perform at the level at which they are capable, teachers should encourage students to bypass difficult items and return to them later to use the available time wisely (Brookhart & Nitko, 2019). Therefore, presenting only one item at a time and not permitting students to return to previous items may produce test scores that do not accurately reflect students' abilities.

Creating and Using Different Forms of an Examination With the Same Group of Students

As discussed in Chapter 2, Qualities of Effective Assessment Procedures: Validity, Reliability, and Usability, alternate forms of a test are considered to be equivalent if they were developed from the same test blueprint or table of specifications, and if they produce highly correlated results. Equivalent test forms are widely used in standardized testing to ensure test security, but alternate forms of teacher-constructed tests usually are not subjected to the rigorous process of obtaining empirical data to document their equivalence. Therefore, alternate forms of a test for the same group of students may produce results that are not comparable, leading to inaccurate interpretations of test scores.

Developing and Administering Open-Book Tests

Tests developed for use in traditional courses usually do not permit test-takers to consult references or other resources to arrive at correct responses, and most academic honesty codes and policies include expectations that students will not consult such resources during assessments without the teacher's permission. However, for online assessments, particularly at the graduate level, teachers may develop tests that permit or encourage students to make use of appropriate resources to select or supply correct answers. Commonly referred to as *open-book* or *take-home* tests, these assessments should gauge students' higher order thinking abilities by requiring use of knowledge and skill in novel situations. One of the higher order skills that may be important to assess is the ability to identify and use appropriate reference materials for problem solving, decision making, and clinical reasoning. Teachers can use test item formats such as essay and context-dependent item sets (interpretive exercises) to craft novel materials for students to analyze, synthesize, and evaluate. Because these item formats typically require more time than true-false, multiple-choice, matching, and completion items, teachers should allot sufficient time for online open-book testing. Therefore, administering an open-book assessment as a tightly timed examination to deter cheating will not only produce results that do not accurately reflect students' true abilities but will likely also engender unproductive feelings of anxiety and anger among students (Brookhart & Nitko, 2019).

An additional low-technology strategy used to deter cheating may be the administration of tests in a timed synchronous manner, in which students' test results are not revealed until after all students have finished the examination. Although synchronous online testing may be inconvenient, adequate advance knowledge of test days and times could alleviate scheduling conflicts that some students may encounter.

High-technology solutions to prevent cheating on unproctored tests include browser security programs such as Respondus™ to keep students from searching the Internet while taking the examination (Hart & Morgan, 2009). However, this security feature does not prevent students from using a second computer or seeking assistance from other people during the test. For those wanting to use the best technology available to prevent academic dishonesty, faculty members could use remote proctoring to ensure student identity and monitor student actions (Dunn, Meine, & McCarley, 2010). Remote proctoring, also called *virtual proctoring*, is a service that works within an online test delivery system to emulate the role of an on-site proctor, confirming the identity of the test-taker and safeguarding the integrity of the exam. Proctoring may be synchronous (in real time) or asynchronous (recording the testing session for later review by a proctor), and may be performed by a live person or by automated monitoring technology. Remote proctoring systems usually incorporate devices such as web cameras, microphones, and even biometric scanners into the learning management system.

Students also may be required to use webcams to confirm their identities to the faculty member. Some course management systems have password-protected access and codes to prevent printing, copying, and pasting. Additional antichecking methods include requiring an online password that is different for each test and changing log-in codes just prior to testing. However, these methods do not prevent students from receiving help from other students. Therefore, a reasonable compromise to these dilemmas may be the use of proctored testing centers (Stonecipher & Wilson, 2014).

Many universities and colleges around the country cooperate to offer students the opportunity to take proctored examinations close to their homes. Proctors should be approved by the faculty in advance to observe students taking the examination online (Hart & Morgan, 2009) and should sign an agreement to keep all test materials secure and maintain confidentiality. Although the administration of proctored examinations is not as convenient as an asynchronous nonproctored test, it offers a greater level of assurance that students are taking examinations independently.

Course Assignments

Course assignments may require adjustment for online learning to suit the electronic medium. Online course assignments can be crafted to provide opportunities for students to develop and demonstrate cognitive, affective, and psychomotor abilities. Table 11.1 provides specific examples of learning products in the cognitive, affective, and psychomotor domains that can be used for formative and summative

TABLE 11.1 Examples of Learning Products Used for Online Assessment of Learning

COGNITIVE DOMAIN	AFFECTIVE DOMAIN	PSYCHOMOTOR DOMAIN
Discussion boards Online chats Case analysis Term papers Research or evidence-based practice papers Short written assignments Journals Electronic portfolios	Discussion boards Online chats Case analysis Debates Role-play Discussions of ethical issues Interviews Journals Blogs	Video recordings, use of other technologies Virtual simulations Developing web pages Web page presentations Interactive modules Presentations

evaluation. Assignments such as analyses of cases and critical thinking vignettes, discussion boards, and classroom assessment techniques may be used for formative evaluation, whereas papers, debates, electronic presentations, portfolios, and tests are more frequently used to provide information for summative evaluation (O'Neill, Fisher, & Newbold, 2009). Online course assignments may be used for formative or summative evaluation of student learning outcomes. However, the teacher should make it clear to the students how the assignments are being used for evaluation. No matter what type of assignment the faculty member assesses, the student must have clearly defined criteria for the assignment and its evaluation.

Feedback

As in traditional courses, feedback during the learning process and following teacher evaluation of assignments facilitates learning. Students need more feedback in online learning than in the traditional environment because of the lack of face-to-face interaction and subsequent lack of nonverbal communication. Teachers should give timely feedback about each assignment to verify that they are in the process of or have finished assessing it, or to inform the student when to expect more detailed feedback. O'Neill et al. (2009) suggested that feedback should be given within 24 to 48 hours, but it may not be reasonable to expect teachers to give detailed, meaningful feedback to a large group of students or on a lengthy assignment within that time frame. For this reason, the syllabus for an online or a hybrid course should include information about reasonable expectations regarding the timing of feedback from the teacher. For example, the syllabus might state, "I will acknowledge receipt of submitted assignments via e-mail within 24 hours, and I will e-mail [or post as a private message on the learning management system, or other means] more detailed, specific feedback [along with a score or grade if appropriate] within [specify time frame]."

Feedback to students can occur through a variety of methods. Many faculty members provide electronic feedback on written assignments using the Track Changes

feature of Microsoft Word (or similar feature of other word-processing software) or by inserting comments into the document. Feedback also may occur through email, orally using vodcasting or scheduled phone conferences, or via a telecommunications application such as Skype, FaceTime, or Zoom.

As discussed in Chapter 9, Assessment of Written Assignments, the teacher may also incorporate peer critique within the process of completing an assignment. For example, for a lengthy written formal paper, the teacher may assign each student a peer-review partner, or each student may ask a peer to critique an early draft. The peer reviewer's written feedback and the resulting revision should then be submitted for the faculty member to assess.

When an assignment involves participation in discussion using the course management system's discussion forum, the teacher may also assign groups or partners to critique each other's posted responses to questions posed by the teacher or other students. Although peer feedback is important to identify areas in which a student's discussion contribution is unclear or incomplete, the course faculty member should also post summarized feedback to the student group periodically to identify gaps in knowledge, correct misinformation, and help students construct new knowledge.

No matter which types of feedback a teacher chooses to use in an online course, clear guidelines and expectations should be established and clearly communicated to the learners, including due dates for peer feedback. Students should understand the overall purpose of feedback to effectively engage in these processes. Structured feedback forms may be used for individual or group work. O'Neill et al. (2009) recommended multidimensional feedback that:

- Addresses the content of the assignment, quality of the presentation, and grammar and other technical writing qualities.
- Provides supportive statements highlighting the strengths and areas of improvement.
- Conveys a clear, thorough, consistent, equitable, constructive, and professional message.

Development of a scoring rubric provides an assessment tool that uses clearly defined criteria for the assignment and gauges student achievement. Rubrics enhance assessment reliability among multiple graders, communicate specific goals to students, describe behaviors that constitute a specific grade, and serve as a feedback tool. Table 11.2 provides a sample rubric for feedback about an online discussion board assignment.

Assessing Student Clinical Performance

Clinical evaluation of students in online courses and programs presents challenges to faculty members and program administrators. When using an online delivery mode, it is critical to ensure the clinical competence of nursing students. Although

TABLE 11.2 Example of Discussion Board Feedback Rubric

CRITERIA	EXEMPLARY (3 POINTS)	GOOD (2 POINTS)	SATISFACTORY (1 POINT)	UNSATISFACTORY (0 POINTS)	SCORE
Frequency	Participates four to five times during a week	Participates two to three times during the week	Participates once near the end of the week	No participation on discussion board	
Initial assignment posting	Posts a well-developed discussion that addresses three or more concepts related to the topic	Posts a well-developed discussion addressing at least one or two key concepts related to the topic	Posts a summary with superficial preparation and unsupported discussion	No assignment posted	
Peer feedback postings	Posts an analysis of a peer's post extending the discussion with supporting references	Posts a response that elaborates on a peer's comments with references	Posts superficial responses such as "I agree" or "great idea"	Does not post feedback to peers	
Content	Post provides a reflective contribution with evidence-based references extending the discussion	Post provides evidence-based facts supporting the topic	Post does not add substantive information to the discussion	Post does not relate to the topic	
References	Provides personal experiences and reflection with two or more supporting references	Provides personal experiences and only one supporting reference	Provides personal experiences and no references	Provides no personal experience or references	
Grammar, clarity, writing style	Responses organized, no grammatical or spelling errors, correct style	Responses organized, one to two grammatical and spelling errors, uses correct style	Responses organized, three to four grammatical and spelling errors, one to two minor style errors	Responses are not organized, five to six grammatical and spelling errors, many style errors	

the didactic component of nursing courses may lend itself well to online delivery, teaching and evaluating clinical skills can prove more challenging in an online context (Bouchoucha, Wikander, & Wilkin, 2013).

Methods for evaluating student clinical performance in an online course format usually involve one or more of the following approaches:

- Use of preceptors to observe and evaluate performance
- The faculty member travels to the student's location to observe student performance directly
- On-campus or regional evaluation of skills in a simulated setting or with live models or standardized patients
- Virtual site visits using teleconferencing, telehealth technology, video recording, live streaming, or similar technologies (Alton, Luke, & Wilder, 2018; Harris et al., 2020)

Use of Preceptors

Students enrolled in online courses or programs usually work with preceptors for the clinical portion of nursing courses. Preceptors are responsible for guiding the students' learning in the clinical environment according to well-defined learning objectives. They are also responsible for evaluating students by giving them regular feedback about their performance and regularly communicating with faculty regarding students' progress. If students are not able to perform according to expectations, the faculty must be notified so that plans for correcting the deficiencies may be established (Oermann, Shellenbarger, & Gaberson, 2018).

Strategies should be implemented in the course for preceptors and other educators involved in the performance evaluation to discuss as a group the competencies to be rated, what each competency means, and the performance of those competencies at different levels on the rating scale. This critical activity ensures reliability among preceptors and other evaluators. Activities can be provided in which preceptors observe video recordings of performances of students and rate their quality using the clinical evaluation tool. Preceptors and course faculty members then can discuss the performance and rating. Alternately, discussions about levels of performance and their characteristics and how those levels would be reflected in ratings of the performance can be held with preceptors and course faculty members. Preceptor development activities of this type should be done before the course begins and at least once during the course to ensure that evaluators are using the tool as intended and are consistent across student populations and clinical settings. Even in clinical courses involving preceptors, faculty members may decide to evaluate clinical skills themselves by reviewing digital recordings of performance or observing students by using other technology with faculty at the receiving end. Digitally recording performance

is valuable not only as a strategy for summative evaluation, to assess competencies at the end of a clinical course or another designated point in time, but also for review by students for self-assessment and by faculty members to give feedback.

Faculty Observation and Evaluation

Even when preceptors are used to supplement the program faculty, it is the faculty's responsibility to summatively evaluate the student's performance. Many nurse practitioner programs perform on-site evaluations of students in which the faculty member visits the site and observes the interaction of the student with patients and the preceptor (Distler, 2015). Although some students may take both online and face-to-face courses at the same institution, most students enrolled in completely online programs are located at a geographical distance from the offering school. Because of this distance, the time and cost of travel for faculty members to observe each student more than once in the clinical setting during each clinical course may be prohibitive (National Organization of Nurse Practitioner Faculties [NONPF], 2003). Another disadvantage to the on-site evaluation is that the face-to-face faculty and student evaluation can be an uncomfortable time for patients and preceptors (Distler, 2015).

In an issue statement on the clinical evaluation of advanced practice nurse and nurse practitioner students, the National Organization of Nurse Practitioner Faculties reaffirmed the need to "evaluate students cumulatively based on clinical observation of student performance by [nurse practitioner] faculty and the clinical preceptor's assessment" and stated that "[d]irect clinical observation of student performance is essential" (NONPF, 2003). According to the National Task Force on Quality Nurse Practitioner Education (2012), clinical observation may be accomplished using direct or indirect evaluation methods such as student-faculty conferences, computer simulation, videotaped sessions, clinical simulations, or other appropriate technologies.

On-Campus or Regional Evaluation Sites

Many online nursing programs require students to attend an on-campus intensive study and evaluation period yearly or during every academic term. In these settings, the nursing faculty can observe students to determine whether they have achieved a certain level of proficiency. Direct observation often is facilitated through the use of competency assessments such as the Objective Structured Clinical Examination tools (Bouchoucha et al., 2013). Some online programs have designated regional evaluation sites where students can go to have their performance evaluated by a faculty member.

In an on-campus or regional assessment setting, students may be required to demonstrate competency with real patients provided by the student or faculty, or with simulated or standardized patients. A standardized patient is a lay person or actor trained to play the role of a patient with specific needs. Standardized patients have the advantage of training to give specific, immediate feedback to students regarding their skill.

Virtual Site Visits Using Various Technologies

An online mode of course and program delivery affects the faculty's ability to personally verify the assessments that are made of students in geographically distant locations and increases reliance on the preceptor's assessment of the student's performance. Various technologies, such as telehealth technology, teleconferencing, video recording, live streaming, and other technologies, are being used by nursing faculty members to assess students' clinical performance at a distance, conduct virtual site visits similar to a visit done in person in a clinical setting, consult with preceptors, and discuss patients and management approaches with students. With consent from the patients and clinical facilities, students can video record their performance of clinical skills (Strand, Fox-Young, Long, & Bogossian, 2013). An advantage to this technology is that students may view the recording along with their preceptors and faculty members, offering the opportunity to reflect on their own performance and receive feedback.

Clinical Evaluation Methods

The clinical evaluation methods presented in Chapter 14, *Clinical Evaluation Methods*, can be used for evaluation in online courses. The critical decision for the teacher is to identify which clinical competencies and skills, if any, need to be observed and the performance rated because that decision suggests different evaluation methods than if the focus of the evaluation is on the cognitive outcomes of the clinical course. In programs in which students work with preceptors or adjunct faculty available on site, any of the clinical evaluation methods presented in Chapter 14, *Clinical Evaluation Methods*, can be used as long as they are congruent with the course outcomes and competencies to be developed by students. There should be consistency, though, in how the evaluation is done across preceptors.

Simulations and standardized patients are other strategies useful in assessing clinical performance in online courses. Performance with standardized patients can be digitally recorded, and students can submit their patient assessments and other written documentation that would commonly be done in practice in that situation. Students also can complete case analyses related to the standardized patient encounter for assessing their knowledge base and rationale for their decisions. Ballman, Garritano, and Beery (2016) described their use of virtual interactive cases in their distance-based nurse practitioner program. The interactive case is a virtual patient encounter with a standardized patient. The experience is comparable to the student being in an examination room interviewing, collecting data from, and assessing the standardized patient. Students can demonstrate clinical skills and perform procedures on manikins and models, with their performance digitally recorded and transmitted to faculty members for evaluation. In online courses, an e-portfolio is a useful evaluation method because it allows students to provide materials that indicate their achievement of the course outcomes and clinical competencies. Simulations, analyses of cases, case presentations, and written assignments can be used to evaluate students' cognitive skills.

A combination of approaches is more effective than one method alone. Exhibit 11.1 summarizes clinical evaluation methods useful for online nursing clinical courses.

EXHIBIT 11.1

CLINICAL EVALUATION METHODS FOR ONLINE NURSING COURSES

EVALUATION OF PSYCHOMOTOR, PROCEDURAL, AND OTHER CLINICAL SKILLS

Observation of performance (by faculty members on site or at a distance, preceptors, examiners, others):

- With patients; in simulations; in virtual site visits (using varied technologies); with models, manikins, or standardized patients
- Objective Structured Clinical Examinations and standardized patients (in laboratories on site, regional centers, other settings, virtually)

Rating of performance:

- Using rating scales, checklists, performance algorithms
- By faculty members, preceptors, examiners, others on site

Notes about clinical performance by preceptor, examiner, others in local area

EVALUATION OF COGNITIVE OUTCOMES AND SKILLS

Test items on clinical knowledge and higher level cognitive skills

Analyses of clinical situations in own practice, of cases, and of media clips:

- Reported in a paper, in discussion board, as part of other online activities

Written assignments:

- Write-ups of cases, analyses of patient care, and other clinical experiences
- Electronic journals
- Analyses of interactions in clinical setting and simulated experiences
- Written assignments, papers
- Nursing care and management plans
- Sample documentation

Case presentations (can be recorded for faculty members at a distance)

Online conferences, discussions

E-portfolio (with materials documenting clinical competencies developed in practicum)

EVALUATION OF AFFECTIVE OUTCOMES

Online conferences and discussions about values, attitudes, and biases that might influence patient care and decisions; about cultural dimensions of care

Analyses and discussions of cases presented online, of clinical scenarios shown in media clips and other multimedia

Written assignments (e.g., reflective papers, journals, others)

Debates about ethical decisions

Value-clarification strategies

Reflective journals

■ Assessment of Online Courses

Online course assessment involves many of the same criteria used to assess courses offered in traditional classrooms, but additional elements specific to the online environment must also be evaluated, such as technology, accessibility, instructional design, content, and interactive activities (O'Neill et al., 2009). These elements of course evaluation are included in the International Association for K-12 Online Learning (iNACOL) guidelines and recommendations for evaluating online courses (Pape, Wicks, & the iNACOL Quality Standards for Online Programs Committee, 2011). Although developed for elementary and secondary education programs, many, if not all, of the standards also apply to online courses in higher education. Table 11.3 provides a summary of the iNACOL standards. Potential methods for collecting this information include student and teacher end-of-course evaluations, interviews or focus groups with students and teachers (electronically if necessary), and peer evaluation of online courses by other faculty members.

TABLE 11.3 iNACOL National Standards for Quality Online Courses

Content	<p>Course goals or objectives clearly state in measurable terms what the participants will know or be able to do at the end of the course.</p> <p>Course components (objectives or goals, assessments, instructional methods, content, assignments, and technology) are appropriately rigorous.</p> <p>Information literacy skills are integrated into the course.</p> <p>A variety of learning resources and materials are available to students before the course begins (e.g., textbooks, browsers, software, tutorials, orientation).</p> <p>Information is provided to students about how to communicate with the online instructor.</p> <p>A code of conduct, including netiquette standards and expectations for academic integrity, is posted.</p>
Instructional design	<p>Course offers a variety of instructional methods.</p> <p>Course is organized into units or lessons.</p> <p>An overview for each unit or lesson describes objectives, activities, assignments, assessments, and resources.</p> <p>Course activities engage students in active learning (e.g., collaborative learning groups, student-led review sessions, games, concept mapping, case study analysis).</p> <p>A variety of supplemental resources is clearly identified in the course materials.</p>
Student assessment	<p>Methods for assessing student performance or achievement align with course goals or objectives.</p> <p>The course provides frequent or ongoing formative assessments of student learning.</p> <p>Feedback tools are built into the course to allow students to view their progress.</p> <p>Assessment materials provide flexibility to assess students in a variety of ways.</p> <p>Assessment rubrics are provided for each graded assignment.</p>

(continued)

TABLE 11.3 iNACOL National Standards for Quality Online Courses
(continued)

Technology	<p>The course uses consistent navigation methods requiring minimal training.</p> <p>Students can use icons, graphics, and text to move logically through the course.</p> <p>Media are available in multiple formats for ease of access and used to meet diverse student needs (e.g., video, podcast).</p> <p>All technology requirements (hardware, software, browser, etc.) and prerequisite technology skills are specified in the course descriptions before the course begins.</p> <p>The course syllabus clearly states the copyright or licensing status, including permission to share when applicable.</p> <p>Course materials and activities are designed to facilitate access by all students.</p> <p>Student information is protected as required by the Family Educational Rights and Privacy Act.</p>
Course evaluation and support	<p>A combination of students, instructors, content experts, instructional designers, and outside reviewers review the course for effectiveness using multiple data-collection methods (e.g., course evaluations, surveys, peer review).</p> <p>The course is updated annually with the date posted on the course management system and all course documents.</p> <p>The course provider offers technical support and course management assistance to the students and course instructor 24 hours a day, 7 days a week.</p>

Source: Adapted from International Association for K-12 Online Learning. (2011). National standards for quality online courses (version 2). Retrieved from <http://www.inacol.org/wp-content/uploads/2015/02/national-standards-for-quality-online-courses-v2.pdf>

In some ways, online courses are isolated and hidden from the view of faculty members and administrators who are not directly involved in teaching them, limiting the role that these colleagues can play in course evaluation. Unlike courses that are taught in traditional classrooms, faculty peers and administrators cannot walk by an open classroom door for a quick informal observation of activities, easily obtain and review hard copies of student assignments and instructor feedback to the students, or critique a printed copy of a course examination and the test and item analysis that pertains to it. Because course activities may take place within a course management system that controls access to course documents and features such as a discussion board, assignment drop box, and grade book, faculty peer reviewers and administrators must make arrangements to enter the course site to assess elements such as the course design and components, congruence of learning activities with intended course outcomes, availability of learning resources and materials, and ways in which student performance is assessed. Also, if course learning activities are conducted asynchronously, such as posting comments to a discussion board, it is difficult for an outside reviewer to assess such elements as the pace of the learning activities and the timing of instructor feedback to students.

Instructional Design

The design of online courses is an important consideration of course evaluation. Standard 4 in the Council of Regional Accrediting Commissions (C-RAC) guidelines for evaluating online programs (2011), suggests that online course design and delivery methods should facilitate communication and active participation of students with each other and with faculty members (C-RAC, 2011). The C-RAC guidelines for evaluating online programs are discussed in Chapter 18, Program Evaluation and Accreditation. A widely used tool for evaluating the overall course design is the *Quality Matters Higher Education Rubric* (Quality Matters, 2018), which focuses on the alignment of elements of the online course such as learning activities and instructional materials with each other to achieve the learning objectives or competencies.

The design of the *interface* (learning management system) concerns its usability to minimize the cognitive load on students, making the system effective, efficient, and pleasing to users. Although faculty members do not design their learning management systems, they are often asked to serve on a product selection committee. Knowledge of usability can assist the faculty to evaluate and select a learning management system that is user friendly for students and teachers (Frith, 2017).

Teachers can control the design of *navigation* within a learning management system. Most learning management systems allow faculty members to organize their courses in different ways; however, this flexibility can create a barrier to learning if navigation is different from course to course. The faculty of online programs should evaluate the navigation design across all courses in the program and, if necessary, develop navigation templates for their online courses to standardize them (Frith, 2017).

Teachers are content experts in the courses they teach, but they may need to consult with instructional designers to improve the online *delivery of content* to students. Instructional designers who are up to date on the effects of new technologies on pedagogy make excellent partners for faculty content experts. Content delivery methods should be evaluated to determine their effectiveness in increasing interactivity, communication, collaboration, and connection among students (Frith, 2017).

Formative evaluation should be designed into every online course. As students navigate through the content, data about their performance in the course can be generated through quizzes, discussion forums, polls, and other methods. Formative assessments are used to provide feedback to students as they are learning. Students may use this feedback to clarify misunderstood concepts or to gain a deeper understanding of content. In addition, teachers can use feedback from students to take corrective action in the design of the course during its delivery or prior to the next offering (Frith, 2017).

Summative evaluation at the end of a course is performed to assess student learning outcomes and student satisfaction with the course. Other aspects of online courses appropriate for summative evaluation can include technical assistance for students and teachers, support for diverse learning styles, and evaluation of faculty who teach in online courses (Frith, 2017).

■ Assessment of Online Teaching

Many colleges and universities use the same instruments for student evaluation of teaching both in traditional courses and online courses. However, because of the unique features of online courses, including reliance on technology for course delivery, the asynchronous nature of some or all learning activities, and physical separation of teacher and students, additional elements may be added to the student evaluation of teaching to reflect these differences or an entirely different instrument may be used. For example, students in online courses may assess the instructor's skill in using the course management system and other technology, facilitating online discussions and other interactions among students, and responding to student questions and comments within a reasonable period of time.

As with online course assessment, iNACOL standards and guidelines developed for assessing the quality of online teaching in K-12 education (Pape et al., 2011) may be adapted for use in higher education settings, including nursing education programs. Table 11.4 presents iNACOL standards and criteria that may be used to develop instruments for student assessment of online teaching.

TABLE 11.4 iNACOL National Standards for Quality Online Teaching

THE ONLINE INSTRUCTOR:	THE ONLINE INSTRUCTOR PROVIDES A LEARNING ENVIRONMENT THAT ENABLES STUDENTS TO MEET IDENTIFIED LEARNING OUTCOMES BY:
Creates learning activities to enable student success	Using an array of online tools for communication, productivity, collaboration, assessment, presentation, and content delivery Incorporating multimedia and visual resources into online modules
Uses a range of technologies to support student learning	Performing basic troubleshooting skills and addressing basic technical issues of online students
Designs strategies to encourage active learning, interaction, participation, and collaboration in the online environment	Using online instructional strategies based on current research and practice (e.g., discussion, student-directed learning, collaborative learning, lecture, project-based, and collaboration in the learning, discussion forum, group work) Promoting student success through clear expectations, prompt responses, and regular feedback
Guides legal, ethical, and safe behavior related to technology use	Providing "netiquette" guidelines in the syllabus Establishing criteria for appropriate online behavior for both teacher and students
Demonstrates competence in creating and implementing assessments in online learning environments	Updating knowledge and skills of evolving technology that support online students' learning styles Addressing the diversity of student academic needs Recognizing and addressing the inappropriate use of electronically accessed data or information

Source: Adapted from International Association for K-12 Online Learning. (2011). National standards for quality online courses (version 2). Retrieved from <http://www.inacol.org/wp-content/uploads/2015/02/national-standards-for-quality-online-courses-v2.pdf>.

Student Evaluation of Teaching

A common challenge to administering online surveys for student assessment of teaching, however, is a response rate lower than that usually achieved when surveys are distributed to students in traditional courses during a regular class period by someone other than the teacher and without the teacher present. The low response rate may be attributed to students' concerns about whether their responses will be anonymous or whether the teacher will be able to identify the source of specific ratings and comments, especially if surveys are administered within the learning management system. One potential solution is to make electronic student assessment of teaching available from college or university websites that are separate from learning management systems and specific course sites.

Peer Review of Teaching

Peer evaluation of teaching can be conducted for online courses as well as for on-campus settings. By reviewing course materials and visiting course websites as guest users, peer evaluators of teaching in online courses can look for evidence that teachers demonstrate application of the following principles of effective instruction, such as:

- How quickly and thoroughly does the teacher respond to student questions?
- Does the teacher use group assignments, discussion boards, or peer critique of assignments to promote interaction and collaboration among students?
- Does the teacher use assignments that require the active involvement of students in their own learning?
- Does the teacher provide prompt, meaningful feedback on assignments posted to a course website or submitted via email?
- Is there evidence that students are actively engaged and spend an appropriate amount of time on course tasks?
- Does the teacher have realistically high standards for achievement of course objectives and communicate them to students?
- Does the teacher accommodate a variety of learning modes, views, abilities, and preferences?
- Is the online course well organized, with easy-to-locate course material and clear directions?
- Is the web design for the course inviting, are graphics used appropriately, and is color used in an appealing way?

■ Assessing Quality of Online Programs

Assessing the quality of online programs is a formal process of measuring quality indicators, using the data to develop an improvement plan, and reassessing the indicators to determine program effectiveness. A nursing program might include the online assessment plan as part of comprehensive assessment plans for on-campus programs or develop it separately. In either case, nurse educators and program administrators should work together to design an assessment plan that leads to continuous improvement and data-driven decision-making about the online program (Frith, 2017).

Several widely used frameworks for assessing quality in online courses and programs include the Western Interstate Commission for Higher Education's *Principles of Good Practice*, Online Learning Consortium's (formerly Sloan-C) *Quality Framework*, and *Quality Matters* (Billings, Dickerson, Greenberg, Yow-Wu, & Talley, 2013). Other frameworks for evaluating distance education programs are included in Chapter 18, Program Evaluation and Accreditation. These frameworks have many common themes that can assist faculties and program administrators with evaluating and improving the overall quality of their online courses and programs:

- Institutional commitment, support, and leadership
- Teaching and learning
- Faculty support
- Student support
- Institutional support for course development
- Technology
- Evaluation and assessment
- Cost-effectiveness
- Management and planning
- Faculty and student satisfaction

The faculty can adapt a framework for assessing quality in online programs by selecting representative indicators from each part of a framework. Once the framework and indicators are identified, the quality improvement plan can be developed, including benchmarks, data sources, persons responsible for assessment, assessment frequency, actual outcomes, action plan, and action results. The assessment plan then guides faculty and administrators to be deliberate in their approach to quality improvement (Frith, 2017).

■ Summary

This chapter discussed methods of assessing learning in online courses, including testing and course assignments, to determine whether course goals have been met.

Online assessment principles do not differ substantially from the approaches used in the traditional classroom environment, but courses and programs that are offered only online or in a hybrid format depend heavily or entirely on technological methods to assess learning.

The use of online testing usually raises concerns among teachers about academic dishonesty. Faculty members want to be confident that students who are enrolled in the course are the ones who are taking the tests, and they want to prevent students from using unauthorized sources of information during a test or sharing test information with students who have not yet taken it. A number of low- and high-technology solutions have been proposed to deter cheating on online tests. Each of these options has advantages and disadvantages that were discussed in the chapter.

Course assignments usually require some adaptation for online learning. The teacher should make it clear to the students how the assignments are being used for evaluation and clearly define the criteria for each assignment. Students need more feedback during the learning process in online learning than in traditional courses because of the lack of face-to-face interaction. Teachers should give timely feedback about each assignment, and the syllabus for an online or hybrid course should include information about reasonable expectations regarding the timing of feedback from the faculty member. Feedback to students about assignments can be provided through a variety of methods. Scoring rubrics that clearly define criteria for the assignment enhance assessment reliability, communicate specific goals to students, describe what behaviors constitute a specific grade, and serve as a feedback tool.

Clinical evaluation of students in online courses and programs presents a variety of challenges. A number of approaches were discussed, including use of preceptors, direct observation by the faculty member, use of standardized patients, video recording, and using other technologies for conducting virtual site visits.

The chapter also discussed modifications of program assessment approaches for online nursing courses and programs. Standards for assessing these were described.

■ References

- Alton, S., Luke, S. A., & Wilder, M. (2018). Cost-effective virtual clinical site visits for nurse practitioner students. *Journal of Nursing Education*, 57, 308–311. doi:10.3928/01484834-20180420-11
- American Association of Colleges of Nursing. (2019). *Degree completion programs for registered nurses: RN to master's degree and RN to baccalaureate programs*. Washington, DC: Author. Retrieved from <https://www.aacnnursing.org/News-Information/Fact-Sheets/Degree-Completion-Programs>
- Ballman, K., Garritano, N., & Beery, T. (2016). Broadening the reach of standardized patients in nurse practitioner education to include the distance learner. *Nurse Educator*, 41, 230–233. doi:10.1097/NNE.0000000000000260
- Billings, D. M., Dickerson, S., Greenberg, M., Yow-Wu, B., & Talley, B. (2013). Quality monitoring and accreditation in nursing distance education programs. In K. Frith & D. Clark (Eds.), *Distance education in nursing* (3rd ed.). New York, NY: Springer Publishing Company.

- Bouchoucha, S., Wikander, L., & Wilkin, C. (2013). Assessment of simulated clinical skills and distance students: Can we do it better? *Nurse Education Today*, 33, 944–948.
- Brookhart, S. M., & Nitko, A. J. (2019). *Educational assessment of students* (8th ed.). New York, NY: Pearson Education.
- Conway-Klaassen, J., & Keil, D. (2010). Discouraging academic dishonesty in online courses. *Clinical Laboratory Science*, 23, 194–200.
- Council of Regional Accrediting Commissions. (2011). *Distance education programs: Interregional guidelines for the evaluation of distance education*. Philadelphia, PA: Middle States Commission on Higher Education. Retrieved from www.metro.inter.edu/distancia/documentos/Guidelines-for-the-Evaluation0of-Distance-Education-Programs.pdf
- Distler, J. W. (2015). Online nurse practitioner education: Achieving student competencies. *The Nurse Practitioner*, 40(11), 44–49. doi:10.1097/01.NPR.0000472249.05833.49
- Dunn, T. P., Meine, M. F., & McCarley, J. (2010). The remote proctor: An innovative technological solution for online course integrity. *International Journal of Technology, Knowledge, and Society*, 6(1), 1–7.
- Frith, K. H. (2017). Assessment of online courses and programs. In M. H. Oermann (Ed.), *A systematic approach to assessment and evaluation of nursing programs* (pp. 103–117). Philadelphia, PA: Wolters Kluwer/National League for Nursing.
- Gronlund, N. E. (2006). *Assessment of student achievement* (8th ed.). Boston, MA: Allyn & Bacon.
- Harris, M., Rhoads, S. J., Rooker, J. S., Kelly, M. A., Lefler, L., Lubin, S., ... Beverly, C. J. (in press). Using virtual site visits in the clinical evaluation of nurse practitioner students: Student and faculty perspectives. *Nurse Educator*. In press. doi:10.1097/nne.0000000000000693
- Hart, L., & Morgan, L. (2009). Strategies for online test security. *Nurse Educator*, 34, 249–253.
- International Association for K-12 Online Learning. (2011). *National standards for quality online courses (version 2)*. Retrieved from <http://www.inacol.org/wp-content/uploads/2015/02/national-standards-for-quality-online-courses-v2.pdf>
- Jones, D., & Wolf, D. (2010). Shaping the future of nursing education today using distance education and technology. *ABNF Journal*, 21(2), 44–47.
- Miller, M. D., Linn, R. L., & Gronlund, N. E. (2013). *Measurement and assessment in teaching* (11th ed.). Upper Saddle River, NJ: Prentice Hall.
- Morgan, L., & Hart, L. (2013). Promoting academic integrity in an online RN–BSN program. *Nursing Education Perspectives*, 34, 240–243.
- National Organization of Nurse Practitioner Faculties. (2003). *NONPF issue statement on clinical evaluation of APN/NP students*. Washington, DC: Author. Retrieved from <http://c.ymcdn.com/sites/www.nonpf.org/resource/resmgr/imported/clinobserv2003.pdf>
- National Task Force on Quality Nurse Practitioner Education. (2012). *Criteria for evaluation of nurse practitioner programs* (4th ed.). Washington, DC: National Organization of Nurse Practitioner Faculties. Retrieved from <http://c.ymcdn.com/sites/www.nonpf.org/resource/resmgr/docs/ntfevalcriteria2012final.pdf>
- Oermann, M. H., Shellenbarger, T., & Gaberson, K. B. (2018). *Clinical teaching strategies in nursing* (5th ed.). New York, NY: Springer Publishing Company.
- O'Neill, C. A., Fisher, C. A., & Newbold, S. K. (2009). *Developing an online course: Best practices for nurse educators* (2nd ed.). New York, NY: Springer Publishing Company.
- Pape, L., Wicks, M., & the iNACOL Quality Standards for Online Programs Committee. (2011). *National standards for quality online programs*. Vienna, VA: International Association for K-12 Online Learning. Retrieved from <http://www.eric.ed.gov/PDFS/ED509638.pdf>

- Quality Matters. (2018). *Higher education rubric* (6th ed.). Annapolis, MD: Author. Retrieved from <https://www.qualitymatters.org/qa-resources/rubric-standards/higher-ed-rubric>
- Seaman, J. E., Allen, I. E., & Seaman, J. (2018). *Grade increase: Tracking online education in the United States*. Babson Park, MA: Babson Survey Research Group. Retrieved from <http://onlinelearningsurvey.com/reports/gradeincrease.pdf>
- Stonecypher, K., & Wilson, P. (2014). Academic policies and practices to deter cheating in nursing education. *Nursing Education Perspectives*, 35, 167–179.
- Strand, H., Fox-Young, S., Long, P., & Bogossian, F. (2013). A pilot project in distance education: Nurse practitioner students' experience of personal video capture technology as an assessment method of clinical skills. *Nurse Education Today*, 33, 253–257.
- Wideman, M. (2011). Caring or collusion? Academic dishonesty in a school of nursing. *Canadian Journal of Higher Education*, 41(2), 28–43.

TEST AND ITEM ANALYSIS: INTERPRETING TEST RESULTS

After a test is scored, the teacher needs to interpret the results and use these interpretations to make grading, selection, placement, or other decisions. To accurately interpret test scores, the teacher needs to analyze the performance of the test as a whole and of the individual test items, and to use these data to draw valid inferences about student performance. This information also helps teachers prepare for posttest discussions with students about the exam. This chapter discusses the process of performing test and item analyses. It also suggests ways in which teachers can use posttest discussions to contribute to student learning and seek student feedback that can lead to test-item improvement.

■ Interpreting Test Scores

As a measurement tool, a test results in a score—a number. A number, however, has no intrinsic meaning and must be compared with something that has meaning to interpret its significance. For a test score to be useful for making decisions about the test, the teacher must interpret the score. Whether the interpretations are norm referenced or criterion referenced, a basic knowledge of statistical concepts is necessary to assess the quality of tests (whether teacher-made or published), understand standardized test scores, summarize assessment results, and explain test scores to others.

Test Score Distributions

Some information about how a test performed as a measurement instrument can be obtained from computer-generated test- and item-analysis reports. In addition to providing item-analysis data such as difficulty and discrimination indexes, such reports often summarize the characteristics of the score distribution. If the teacher does not have access to electronic scoring and computer software for test and item analysis, many of these analyses can be done by hand, albeit more slowly.

When a test is scored, the teacher is left with a collection of raw scores. Often these scores are recorded according to the names of the students, in alphabetical order, or by student numbers. As an example, suppose that the scores displayed in Table 12.1 resulted from the administration of a 65-point test to 16 nursing students. Glancing at this collection of numbers, the teacher would find it difficult to answer such questions as:

- 1. Did a majority of students obtain high or low scores on the test?
 - 2. Did any individuals score much higher or much lower than the majority of the students?
 - 3. Are the scores widely scattered or grouped together?
 - 4. What was the range of scores obtained by the majority of the students?
- (Brookhart & Nitko, 2019)

To make it easier to see similar characteristics of scores, the teacher should arrange them in rank order, from highest to lowest (Miller, Linn, & Gronlund, 2013), as in Table 12.2. Ordering the scores in this way makes it obvious that they ranged from 42 to 60, and that one student’s score was much lower than those of the other students. But the teacher still cannot visualize easily how a typical student performed on the test or the general characteristics of the obtained scores. Removing student names, listing each score once, and tallying how many times each score occurs results in a frequency distribution, as in Table 12.3. By displaying scores in this way, it is easier for the teacher to identify how well the group of students performed on the exam.

TABLE 12.1 List of Students in a Class and Their Raw Scores on a 65-Point Test			
STUDENT	SCORE	STUDENT	SCORE
A. Allen	53	I. Ignatius	48
B. Brown	54	J. Jimanez	55
C. Chen	52	K. Kelly	52
D. Dunlap	52	L. Lynch	42
E. Edwards	54	M. Meyer	47
F. Finley	57	N. Nardozzi	60
G. Gunther	54	O. O’Malley	55
H. Hernandez	56	P. Purdy	53

TABLE 12.2 Rank Order of Students From Table 12.1 With Raw Scores Ordered From Highest to Lowest

STUDENT	SCORE	STUDENT	SCORE
N. Nardozzi	60	A. Allen	53
F. Finley	57	P. Purdy	53
H. Hernandez	56	C. Chen	52
J. Jimanez	55	K. Kelly	52
O. O'Malley	55	D. Dunlap	52
B. Brown	54	I. Ignatius	48
E. Edwards	54	M. Meyer	47
G. Gunther	54	L. Lynch	42

TABLE 12.3 Frequency Distribution of Raw Scores From Table 12.1

RAW SCORE	FREQUENCY
61	0
60	1
59	0
58	0
57	1
56	1
55	2
54	3
53	2
52	3
51	0
50	0
49	0
48	1
47	1
46	0
45	0
44	0
43	0
42	1
41	0

The frequency distribution also can be represented graphically as a histogram. In Figure 12.1, the scores are ordered from lowest to highest along a horizontal line, left to right, and the number of asterisks above each score indicates the frequency of that score. Frequencies also can be indicated on a histogram by bars, with the height of each bar representing the frequency of the corresponding score, as in Figure 12.2.

A frequency polygon is another way to display a score distribution graphically. A dot is made above each score value to indicate the frequency with which that score occurred; if no one obtained a particular score, the dot is made on the base-line, at 0. The dots then are connected with straight lines to form a polygon or curve. Figure 12.3 shows a frequency polygon based on the histogram in Figure 12.1. Histograms and frequency polygons thus show general characteristics such as the scores that occurred most frequently, the score distribution shape, and the range of the scores.

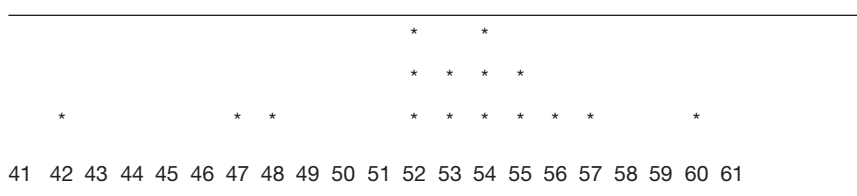


FIGURE 12.1 Histogram depicting frequency distribution of raw scores from Table 12.1.

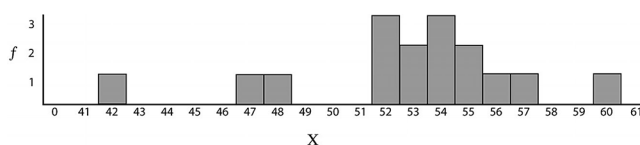


FIGURE 12.2 Bar graph depicting frequency distribution of raw scores from Table 12.1.

Note: f , frequency; X , scores.

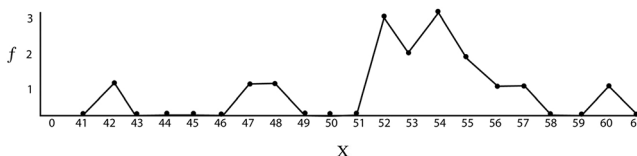


FIGURE 12.3 Frequency polygon depicting frequency distribution of raw scores from Table 12.1.

Note: f , frequency; X , scores.

The characteristics of a score distribution can be described on the basis of its symmetry, skewness, modality, and kurtosis. These characteristics are illustrated in Figure 12.4. A symmetric distribution or curve is one in which there are two equal halves, mirror images of each other. Nonsymmetric or asymmetric curves have a cluster of scores or a peak at one end and a tail extending toward the other end. This type of curve is said to be skewed; the direction in which the tail extends indicates whether the distribution is positively or negatively skewed. The tail of a positively skewed curve extends toward the right, in the direction of positive numbers on a scale, and the tail of a negatively skewed curve extends toward the left, in the direction of negative numbers. A positively skewed distribution, thus, has the largest cluster of scores at the low end of the distribution, which seems counterintuitive. The distribution of test scores from Table 12.1 is nonsymmetric and negatively skewed. Remember that the lowest possible score on this test was 0 and the highest possible score was 65; the scores were clustered between 42 and 60.

Frequency polygons and histograms can differ in the number of peaks they contain; this characteristic is called *modality*, referring to the mode or the most frequently occurring score in the distribution. If a curve has one peak, it is unimodal; if it contains two peaks, it is bimodal. A curve with many peaks is multimodal. The relative flatness or peakedness of the curve is referred to as *kurtosis*. Flat curves are described as *platykurtic*, moderate curves are said to be *mesokurtic*, and sharply peaked curves are referred to as *leptokurtic* (Waltz, Strickland, & Lenz, 2010). The histogram in Figure 12.1 shows a bimodal, platykurtic distribution.

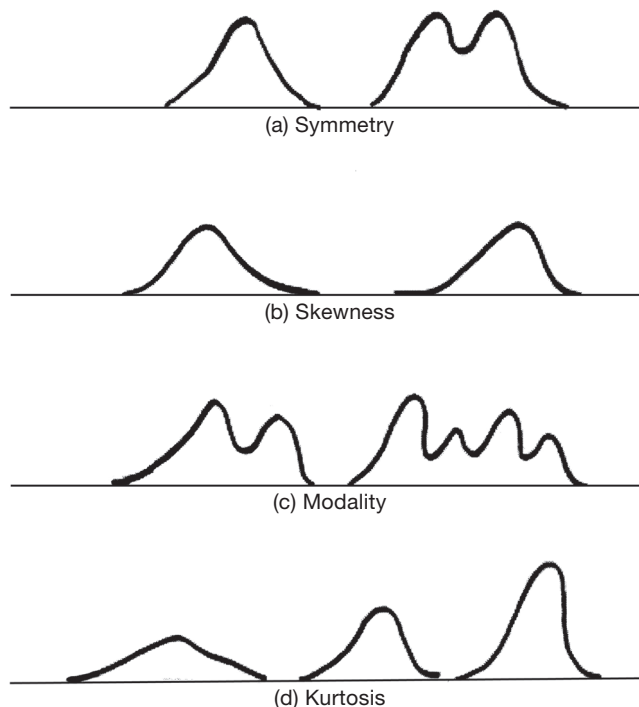


FIGURE 12.4 Characteristics of a score distribution.

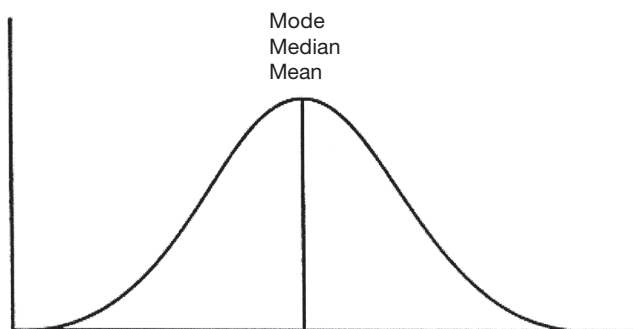


FIGURE 12.5 The normal distribution.

The shape of a score distribution depends on the characteristics of the test as well as the abilities of the students who were tested (Brookhart & Nitko, 2019). Some teachers make grading decisions as if all test score distributions resemble a normal curve, that is, they attempt to “curve” the grades. An understanding of the characteristics of a normal curve would dispel this notion. A normal distribution is a bell-shaped curve that is symmetric, unimodal, and mesokurtic. Figure 12.5 illustrates a normal distribution.

Many human characteristics, such as intelligence, weight, and height, are normally distributed; the measurement of any of these attributes in a population would result in more scores in the middle range than at either extreme. However, most score distributions obtained from teacher-made tests do not approximate a normal distribution. This is true for several reasons. The characteristics of a test greatly influence the resulting score distribution; a very difficult test tends to yield a positively skewed curve. Likewise, the abilities of the students influence the test score distribution. Regardless of the distribution of the attribute of intelligence among the human population, this characteristic is not likely to be distributed normally among a class of nursing students or a group of newly hired RNs. Because admission and hiring decisions tend to select those individuals who are most likely to succeed in the nursing program or job, a distribution of IQ scores from a class of 16 nursing students or 16 newly hired RNs would tend to be negatively skewed. Likewise, knowledge of nursing content is not likely to be normally distributed because those who have been admitted to a nursing education program or hired as staff nurses are not representative of the population in general. Therefore, grading procedures that attempt to apply the characteristics of the normal curve to a test score distribution are likely to result in unwise and unfair decisions.

Measures of Central Tendency

One of the questions to be answered when interpreting test scores is, “What score is most characteristic or typical of this distribution?” A typical score is likely to be in the middle of a distribution with the other scores clustered around it; measures of central

tendency provide a value around which the test scores cluster. Three measures of central tendency commonly used to interpret test scores are the mode, median, and mean.

The mode, sometimes abbreviated *Mo*, is the most frequently occurring score in the distribution; it must be a score actually obtained by a student. It can be identified easily from a frequency distribution or graphic display without mathematical calculation. As such, it provides a rough indication of central tendency. The mode, however, is the least stable measure of central tendency because it tends to fluctuate considerably from one sample to another drawn from the same population (Miller et al., 2013). That is, if the same 65-item test that yielded the scores in Table 12.1 were administered to a different group of 16 nursing students in the same program who had taken the same course, the mode might differ considerably. In addition, as in the distribution depicted in Figure 12.1, the mode has two or more values in some distributions, making it difficult to specify one typical score. A uniform distribution of scores has no mode; such distributions are likely to be obtained when the number of students is small, the range of scores is large, and each score is obtained by only one student.

The median (abbreviated *Mdn* or P^{50}) is the point that divides the distribution of scores into equal halves (Miller et al., 2013). It is a value above which fall 50% of the scores and below which fall 50% of the scores; thus, it represents the 50th percentile. The median does not have to be an actual obtained score. In an even number of scores, the median is located halfway between the two middle scores; in an odd number of scores, the median is the middle score. Because the median is an index of location, it is not influenced by the value of each score in the distribution. Thus, it is usually a good indication of a typical score in a skewed distribution containing extremely high or low scores (Miller et al., 2013).

The mean often is referred to as the “average” score in a distribution, reflecting the mathematical calculation that determines this measure of central tendency. It is usually abbreviated as *M* or \bar{X} . The mean is computed by summing each individual score and dividing by the total number of scores, as in the following formula:

$$M = \frac{\sum X}{N} \quad (12.1)$$

where *M* is the mean, $\sum X$ is the sum of the individual scores, and *N* is the total number of scores. Thus, the value of the mean is affected by every score in the distribution (Miller et al., 2013). This property makes it the preferred index of central tendency when a measure of the total distribution is desired. However, the mean is sensitive to the influence of extremely high or low scores in the distribution, and, as such, it may not reflect the typical performance of a group of students.

There is a relationship between the shape of a score distribution and the relative locations of these measures of central tendency. In a normal distribution, the mean, median, and mode have the same value, as shown in Figure 12.5. In a positively

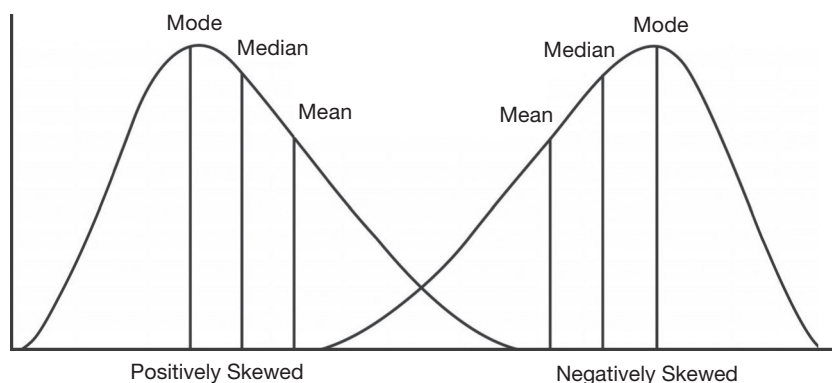


FIGURE 12.6 Measures of central tendency in positively and negatively skewed distributions.

skewed distribution, the mean will yield the highest measure of central tendency and the mode will give the lowest; in a negatively skewed distribution, the mode will be the highest value and the mean the lowest. Figure 12.6 depicts the relative positions of the three measures of central tendency in skewed distributions.

The mean of the distribution of scores from Table 12.1 is 52.75; the median is 53.5. The fact that the median is slightly higher than the mean confirms that the median is an index of location or position and is insensitive to the actual score values in the distribution. The mean, because it is affected by every score in the distribution, was influenced by the one extreme low score. Because the shape of this score distribution was negatively skewed, it is expected that the median would be higher than the mean because the mean is always pulled in the direction of the tail.

Measures of Variability

It is possible for two score distributions to have similar measures of central tendency and yet be very different. The scores in one distribution may be tightly clustered around the mean, and in the other distribution, the scores may be widely dispersed over a range of values. Measures of variability are used to determine how similar or different the students are with respect to their scores on a test.

The simplest measure of variability is the range, the difference between the highest and lowest scores in the distribution. For the test score distribution in Table 12.3, the range is 18 ($60 - 42 = 18$). The range is sometimes expressed as the highest and lowest scores, rather than a difference score. Because the range is based on only two values, it can be highly unstable. The range also tends to increase with sample size; that is, test scores from a large group of students are likely to be scattered over a wide range because of the likelihood that an extreme score will be obtained (Miller et al., 2013).

The standard deviation (abbreviated as *SD*, *s*, or σ) is the most common and useful measure of variability. Like the mean, it takes into consideration every score in the distribution. The standard deviation is based on differences between each score and

the mean. Thus, it characterizes the average amount by which the scores differ from the mean. The standard deviation is calculated in four steps:

1. Subtract the mean from each score ($X - M$) to compute a deviation score (x), which can be positive or negative.
2. Square each deviation score (x^2), which eliminates any negative values. Sum all of the squared deviation scores ($\sum x^2$).
3. Divide this sum by the number of test scores to yield the variance.
4. Calculate the square root of the variance.

Although other formulas can be used to calculate the standard deviation, the following definitional formula represents these four steps:

$$SD = \sqrt{\frac{\sum x^2}{N}} \quad (12.2)$$

where SD is the standard deviation, $\sum x^2$ is the sum of the squared deviation scores, and N is the number of scores (Miller et al., 2013).

The standard deviation of the distribution of scores from Table 12.1 is 4.1. What does this value mean? A standard deviation of 4.1 represents the average deviation of scores from the mean. On a 65-point test, 4 points is not a large average difference in scores. If the scores cluster tightly around the mean, the standard deviation will be a relatively small number; if they are widely scattered over a large range of scores, the standard deviation will be a larger number.

Other Test Characteristics

In addition to interpreting the test score distribution and measures of central tendency and variability, teachers should examine test items in the aggregate for evidence of bias. For example, although there may be no obvious gender bias in any single test item, such a bias may be apparent when all items are reviewed as a group. Similar cases of ethnic, racial, religious, and cultural bias may be found when items are grouped and examined together. The effect of bias on testing and evaluation is discussed in detail in Chapter 16, Social, Ethical, and Legal Issues.

■ Interpreting an Individual Score

Interpreting the Results of Teacher-Made Tests

The ability to interpret the characteristics of a distribution of scores will assist the teacher to make norm-referenced interpretations of the meaning of any individual score in that distribution. For example, how should the teacher interpret P. Purdy's score of 53 on the test whose results were summarized in Table 12.1? With a median

of 53.5, a mean of 52.75, and a standard deviation of 4.1, a score of 53 is about “average.” All scores between 49 and 57 fall within one standard deviation of the mean, and thus are not significantly different from one another. On the other hand, N. Nardozzi can rejoice because a score of 60 is almost two standard deviations higher than the mean; thus, this score represents achievement that is much better than that of others in the group. The teacher should probably plan to counsel L. Lynch, because a score of 42 is more than two standard deviations below the mean, much lower than others in the group.

However, most nurse educators need to make criterion-referenced interpretations of individual test scores. A student’s score on the test is compared with a pre-set standard or criterion, and the scores of the other students are not considered. The percentage-correct score is a derived score that is often used to report the results of tests that are intended for criterion-referenced interpretation. The percentage correct is a comparison of a student’s score with the maximum possible score; it is calculated by dividing the raw score by the total number of items on the test (Miller et al., 2013). Although many teachers believe that percentage-correct scores are an objective indication of how much students really know about a subject, in fact they can change significantly with the difficulty of the test items. Because percentage-correct scores are often used as a basis for assigning letter grades according to a predetermined grading system, it is important to recognize that they are determined more by test difficulty than by true quality of performance. For tests that are more difficult than they were expected to be, the teacher may want to adjust the raw scores before calculating the percentage correct on that test.

The percentage-correct score should not be confused with percentile rank, often used to report the results of standardized tests. The percentile rank describes the student’s relative standing within a group and therefore is a norm-referenced interpretation. The percentile rank of a given raw score is the percentage of scores in the distribution that occur at or below that score. A percentile rank of 83, therefore, means that the student’s score is equal to or higher than the scores made by 83% of the students in that group; one cannot assume, however, that the student answered 83% of the test items correctly. Because there are 99 points that divide a distribution into 100 groups of equal size, the highest percentile rank that can be obtained is the 99th. The median is at the 50th percentile. Differences between percentile ranks mean more at the highest and lowest extremes than they do near the median.

Interpreting the Results of Standardized Tests

The results of standardized tests usually are intended to be used to make norm-referenced interpretations. Before making such interpretations, the teacher should keep in mind that standardized tests are more relevant to general rather than specific instructional goals. In addition, the results of standardized tests are more

appropriate for evaluations of groups rather than individuals. Consequently, standardized test scores should not be used to determine grades for a specific course or to make a decision to hire, promote, or terminate an employee. Like most educational measures, standardized tests provide gross, not precise, data about achievement. Actual differences in performance and achievement are reflected in large score differences.

Standardized test results usually are reported in derived scores such as percentile ranks, standard scores, and norm group scores. Because all of these derived scores should be interpreted in a norm-referenced way, it is important to specify an appropriate norm group for comparison. The user's manual for any standardized test typically presents norm tables in which each raw score is matched with an equivalent derived score. Standardized test manuals may contain a number of norm tables; the norm group on which each table is based should be fully described. The teacher should take care to select the norm group that most closely matches the group whose scores will be compared to it (Miller et al., 2013). For example, when interpreting the results of standardized tests in nursing, the performance of a group of baccalaureate nursing students should be compared with a norm group of baccalaureate nursing students. Norm tables sometimes permit finer distinctions such as size of program, geographical region, and public versus private affiliation.

■ Item Analysis

In addition to test statistics, teachers also should examine indicators of performance quality for each item on the exam. When used together, multiple data points—difficulty index, discrimination index, and point biserial correlation coefficient—provide a rich source of information about the performance quality of test items (Ernie, n.d.). However, teachers should not depend solely on these statistical data to judge the quality of exam items. Decisions about individual test items should be made in the context of the content and structure of the item, the teacher's expectations about how the items would perform, and an accurate interpretation of the item statistics (Brookhart & Nitko, 2019).

Computer software for item analysis is widely available for use with electronic answer sheet scanning equipment. Commercially available computer testing applications usually also provide services that produce user reports of item-analysis statistics. For teachers who do not have access to such equipment and software, procedures for analyzing student responses to test items by hand are described in detail later in this section. Regardless of the method used for analysis, teachers should be familiar enough with the meaning of each item-analysis statistic to correctly interpret the results.

Exhibit 12.1 offers an example of a computer-generated item-analysis report. This example lists only the item-analysis data for each of the exam items, without also including the wording of the items and any codes that the teacher may have used to

EXHIBIT 12.1**SAMPLE COMPUTER-GENERATED ITEM-ANALYSIS REPORT**

ITEM STATISTICS (N = 68)										
ITEM	KEY	A	B	C	D	E	OMIT	MULTIPLE RESPONSE	DIFF. INDEX	DISCRIM. INDEX
1	A	44	0	24	0	0	0	0	.65	.34
2	B	0	62	4	2	0	0	0	.91	.06
3	A	59	1	4	4	0	0	0	.87	.35
4	C	12	4	51	1	0	0	0	.75	.19
5	E	23	8	0	8	29	0	0	.43	.21
6	D	2	3	17	46	0	0	0	.68	.17

Note: Diff. Index, difficulty index; Discrim. Index, discrimination index.

classify the content of the items (e.g., content domain, cognitive level, client needs). This format is useful for quickly scanning the data to identify potential problems. Later examples will illustrate item-analysis reports that display the item content, any classification codes, and additional statistics related to the performance of each item's answer options.

Difficulty Index

One useful indication of test-item quality is its difficulty. The most commonly employed index of difficulty is the *P*-level, the value of which ranges from 0 to 1.00, indicating the percentage of students who answered the item correctly. A *P*-value of 0 indicates that no one answered the item correctly, and a value of 1.00 indicates that every student answered the item correctly. A simple formula for calculating the *P*-value is

$$P = \frac{R}{T} \quad (12.3)$$

where *R* is the number of students who responded correctly and *T* is the total number of students who took the test (Brookhart & Nitko, 2019).

The difficulty index commonly is interpreted to mean that items with *P*-values of .20 and below are difficult, and items with *P*-values of .80 and above are easy. However, this interpretation may imply that test items are intrinsically easy or difficult and may not take into account the quality of the instruction or the abilities of the students in

that group. A group of students who were taught by an expert instructor might tend to answer a test item correctly, whereas a group of students with similar abilities who were taught by an ineffectual instructor might tend to answer it incorrectly. Different *P*-values might be produced by students with more or less ability. Thus, test items cannot be classified as easy or difficult without considering how well that content was taught.

The *P*-value also should be interpreted in relationship to the student's probability of guessing the correct response. For example, if all students guess the answer to a true-false item, on the basis of chance alone, the *P*-value of that item should be approximately .50. On a four-option multiple-choice item, chance alone should produce a *P*-value of .25. A four-alternative, multiple-choice item with moderate difficulty therefore would have a *P*-value approximately halfway between chance (.25) and 1.00, or .625. This calculation is explained as follows:

$$1.00 - .25 = .75 \text{ [range of values between .25 and 1.00]}$$

$$\frac{.75}{2} = .375 \text{ [}\frac{1}{2}\text{ of the range of values between .25 and 1.00]}$$

$$.25 + .375 = .625 \text{ [the chance of guessing correctly plus } \frac{1}{2} \text{ of the range of values between that value and 1.00]}$$

For most tests whose results will be interpreted in a norm-referenced way, *P*-values of .30 to .70 for test items are desirable. However, for tests whose results will be interpreted in a criterion-referenced manner, as most tests in nursing education settings are, the difficulty level of test items should be compared between groups (students whose total scores met the criterion and students who did not). If item difficulty levels indicate a relatively easy (*P*-value of .70 or above) or relatively difficult (*P*-value of .30 or below) item, criterion-referenced decisions still will be appropriate if the item correctly classifies students according to the criterion (Waltz et al., 2010).

Very easy and very difficult items have little power to discriminate between students who know the content and students who do not (see “Discrimination Index” in the next section), and they also decrease the reliability of the test scores. Teachers can use item difficulty information to identify the need for remedial work related to specific content or skills, or to identify test items that are ambiguous (Miller et al., 2013).

Another use of the item difficulty index is to compare the performance of upper and lower scoring groups on each item. The first step in this process is to identify the upper and lower groups of students by their total scores on the exam. If the total number of students taking the exam is between 20 and 40, the 10 highest scoring and 10 lowest scoring students are identified. If there are more than 40 students, common practice is to identify the top 27% and the bottom 27% (Brookhart & Nitko, 2019). In a group of 75 exam takers, 20 students each would be in the top 27% and bottom 27% (Ernie, n.d.). The difficulty index for each item is computed separately for each group, and the resulting *P*-values are used to provide a comparative analysis of the

performance of high and low scorers. This comparative analysis is called a *discrimination index*, which is explained next.

Discrimination Index

The discrimination index, D , is a powerful indicator of test-item quality. A positively discriminating item is one that was answered correctly more often by students with high scores on the test than by those whose test scores were low. A negatively discriminating item was answered correctly more often by students with low test scores than by students with high scores. When an equal number of high- and low-scoring students answer the item correctly, the item is nondiscriminating (Brookhart & Nitko, 2019; Miller et al., 2013).

A number of item discrimination indexes are available; a simple method of computing D is

$$D = P_u - P_l \quad (12.4)$$

where P_u is the fraction of students in the high-scoring group who answered the item correctly and P_l is the fraction of students in the low-scoring group who answered the item correctly.

The D -value ranges from -1.00 to $+1.00$. In general, the higher the positive value, the better the test item. An index of $+1.00$ means that all students in the upper group answered correctly, and all students in the lower group answered incorrectly; this indication of maximum positive discriminating power is rarely achieved. D -values of $+.20$ or above are desirable, and the higher the positive value the better. An index of 0 means that equal numbers of students in the upper and lower groups answered the item correctly, and this item has no discriminating power (Miller et al., 2013). Negative D -values signal items that should be reviewed carefully; usually they indicate items that are flawed and need to be revised. One possible interpretation of a negative D -value is that the item was misinterpreted by high scorers or that it provided a clue to low scorers that enabled them to guess the correct answer (Waltz et al., 2010).

When interpreting a D -value, it is important to keep in mind that an item's power to discriminate is highly related to its difficulty index. An item that is answered correctly by all students has a difficulty index of 1.00 ; the discrimination index for this item is 0 , because there is no difference in performance on that item between students whose overall test scores were high and those whose scores were low. Similarly, if all students answered the item incorrectly, the difficulty index is 0 , and the discrimination index is also 0 because there is no discrimination power. Thus, very easy and very difficult items have low discriminating power. Items with a difficulty index of $.50$ make maximum discriminating power possible, but do not guarantee it (Miller et al., 2013).

It is important to keep in mind that item-discriminating power does not indicate item validity. To gather evidence of item validity, the teacher would have to compare each test item to an independent measure of achievement, which is seldom possible for teacher-constructed tests. Standardized tests in the same content area usually measure the achievement of more general objectives, so they are not appropriate as independent criteria. The best measure of the domain of interest usually is the total score on the test if the test has been constructed to correspond to specific instructional objectives and content. Thus, comparing each item's discriminating power to the performance of the entire test determines how effectively each item measures what the entire test measures. However, retaining very easy or very difficult items despite low discriminating power may be desirable so as to measure a representative sample of learning objectives and content (Miller et al., 2013).

Point Biserial Correlation Coefficient

This statistic indicates the correlation between a student's response to an item and his or her overall performance on the exam. This statistic frequently is calculated and provided as part of an item-analysis report from a commercial test development, scoring, and analysis application. The statistic is interpreted in the same manner as the discrimination index previously described. Values range from -1.00 to $+1.00$, with higher positive values indicating that students who performed well on the exam tended to answer the item correctly and lower positive values indicating that students whose overall test performance was poor tended to answer the item incorrectly. A negative point biserial correlation coefficient indicates a negative correlation between the total score and performance on that item; students with low scores tended to answer the item correctly whereas high-scoring students tended to answer incorrectly. As previously discussed, negative correlations may indicate items that are flawed and need to be revised.

Distractor Analysis

Teachers should not make decisions about retaining a test item in its present form, revising it, or eliminating it from future use on the basis of the item statistics alone. Item difficulty and discrimination indexes are not fixed, unchanging characteristics. Item-analysis data for a given test item will vary from one administration to another because of factors such as students' ability levels, quality of instruction, and the size of the group tested. With very small groups of students, if a few students would have changed their responses to the test item, the difficulty and discrimination indexes could change considerably (Miller et al., 2013). Thus, when using these indexes to identify questionable items, the teacher should carefully examine each test item for evidence of poorly functioning distractors, ambiguous alternatives, and miskeying.

Ideally, every distractor should be selected by at least one student in the lower group, and more lower group students than higher group students should select it. A distractor that is not selected by any student in the lower group may contain a technical flaw or may be so implausible as to be obvious even to students who lack knowledge of the correct answer. A distractor may be ambiguous if upper group students tend to choose it with about the same frequency as the keyed, or correct, response. This result usually indicates that there is no single clearly correct or best answer. Poorly functioning and ambiguous distractors may be revised to make them more plausible or to eliminate the ambiguity. If a large number of higher scoring students select a particular incorrect response, the teacher should check to see whether the answer key is correct. However, as previously mentioned, the content of the item along with the statistics should guide the teacher's decision-making (Brookhart & Nitko, 2019; Ermie, n.d.).


Examining Item-Analysis Data in Context

Multiple factors other than the content and structure of the exam item also may affect students' answers to a test item. Supplemental readings may contradict what was discussed in class, students may misinterpret poorly worded distractors, or an item may be miskeyed. Psychometric data cannot identify these factors, hence the need for item analysis accompanied by review of the actual item (Ermie, n.d.). The following examples from a commercial item-analysis report illustrate how to use item statistics and distractor analysis in the context of actual item content to determine whether the items performed as expected and desired. Assume that these sample items were included on a unit exam on nursing of patients with endocrine disorders. The exam contained 85 items, each worth 1 point, and 70 students took the exam.

Item 6 (Exhibit 12.2) was a true-false item, so it is expected that no students chose answer options C, D, and E. The item had a fairly high difficulty index (0.84), meaning that most students (59 of 70) answered correctly. The content of the item suggests that it was testing at a low cognitive level, probably remembering or understanding, so the teacher may be expecting a high difficulty index. However, the discrimination index was low and slightly negative (-0.07). This could mean that one or two students in the high-scoring group answered this item incorrectly. A closer examination of the wording of this item reveals a possible source of the problem: Hyperthyroidism also causes fatigue. This item should be reviewed in the posttest discussion with the goal of determining why students who chose "false" answered the way they did. If high-scoring students who answered incorrectly give that rationale for their answers, the teacher should consider accepting both "true" and "false" as correct answers (see "Conducting Posttest Discussions" and "Eliminating Items or Adding Points" later in this chapter). This item also should be revised before it is used again on another exam.

EXHIBIT 12.2

EXAMPLE OF A TRUE-FALSE ITEM

<div><div>ExamSoft</div><div>QuestionsRubricsAssessmentsCategoriesReportsExam TakersAdmin</div></div>										
QUESTION #	CORRECT RESPONSES			DISC. INDEX	POINT BISERIAL	CORRECT ANSWER	RESPONSE FREQUENCIES (*INDICATES CORRECT ANSWER)			AVG. ANSWER TIME
	DIFF (P)	UPPER	LOWER				TRUE	FALSE	UNANSWERED	
6	0.84	83.97%	84.04%	−0.07	0.20	True	59	11	0	00.08
						Percentage Selected	84.29	15.71	0.0	

Q: T/F Hypothyroidism is manifested by fatigue and lethargy.

*A. True

B. False


Source: Reprinted by permission of ExamSoft Worldwide, Inc., Dallas, TX.

Items 14 to 17 (Exhibit 12.3) comprise a matching exercise. There are three answer options for each item (premise), so it is expected that no students chose answer options D and E; each response option for every item was chosen by at least two students. Items 16 and 17 have fairly high difficulty indexes with positive discrimination indexes near or above +0.20. The teacher may have expected these two items to be relatively easy because they were discussed in class and the students had inpatient clinical experiences with patients receiving those two insulin types. The positive discrimination indexes indicate that higher scoring students tended to answer correctly more often than low-scoring students did. Items 14 and 15 were more difficult, which the teacher may have expected because these insulins were not discussed extensively in class and students did not have as much clinical exposure to patients who were using them. Both items had positive discrimination indexes, although item 14’s discrimination power was very weak. Item 15’s discrimination index of 0.48 was nearly ideal, suggesting that it reliably discriminated between high-scoring and low-scoring students. On the basis of this review, the teacher may decide that this matching exercise does not need to be revised for future use.

Item 31 (Exhibit 12.4) is a four-option multiple-choice item that was moderately difficult ($P = 0.33$) and negatively discriminating ($D = -0.16$). Each distractor was selected by multiple students, but 30 of the 70 students chose the same incorrect option, C. The negative discrimination index suggests that some of those 30 students were among the high scorers on the test. A closer examination of the distractors reveals a possible explanation: lactic acidosis is not a common side effect of

EXHIBIT 12.3

EXAMPLE OF A MATCHING EXERCISE



Questions Rubrics Assessments Categories Reports Exam Takers Admin

QUESTION #	CORRECT RESPONSES			DISC. INDEX	POINT BISERIAL	CORRECT ANSWER	RESPONSE FREQUENCIES (*INDICATES CORRECT ANSWER)						AVG. ANSWER TIME
	DIFF (P)	UPPER	LOWER				A	B	C	D	E	UNANSWERED	
14	0.24	28.11%	23.06%	0.05	0.03	A	*17	51	2	0	0	0	00:34
						Percentage Selected	24.29	72.86	2.86	0.00	0.00	0.0	
						Point Biserial	0.03	0.28	-0.19	0.00	0.00		
						Disc. Index	0.05	0.50	-0.55	0.00	0.00		
						Upper 27%	0.28	0.72	0.00	0.00	0.00		
						Lower 27%	0.23	0.22	0.55	0.00	0.00		

Q: For each type of insulin in Column A, select its peak action from Column B.

Column A

Column B

14. Detemir

a. Long acting

15. Glargine

b. Intermediate acting

16. NPH

c. Short acting

17. Regular

What is peak action for: Detemir

*A. Long acting

B. Intermediate acting

C. Short acting

EXHIBIT 12.3

EXAMPLE OF A MATCHING EXERCISE (continued)

QUESTION #	CORRECT RESPONSES			DISC. INDEX	POINT BISERIAL	CORRECT ANSWER	RESPONSE FREQUENCIES (*INDICATES CORRECT ANSWER)						AVG. ANSWER TIME
	DIFF (P)	UPPER	LOWER				A	B	C	D	E	UNANSWERED	
15	0.33	61.19%	13.33%	0.48	0.36	A	*24	30	16	0	0	0	00.28
						Percentage Selected	34.29	42.86	22.86	0.00	0.00	0.0	
						Point Biserial	0.36	0.02	-0.21	0.00	0.00		
						Disc. Index	0.48	-0.25	-0.23	0.00	0.00		
						Upper 27%	0.61	0.33	0.06	0.00	0.00		
						Lower 27%	0.13	0.58	0.29	0.00	0.00		

Q: For each type of insulin in Column A, select its peak action from Column B.

- Column A

14. Detemir

15. Glargine

16. NPH

17. Regular
- Column B

a. Long acting

b. Intermediate acting

c. Short acting

What is peak action for: Glargine

- *A. Long acting
- B. Intermediate acting
- C. Short acting

(continued)

EXHIBIT 12.3

EXAMPLE OF A MATCHING EXERCISE (*continued*)

QUESTION #	CORRECT RESPONSES			DISC. INDEX	POINT BISERIAL	CORRECT ANSWER	RESPONSE FREQUENCIES (*INDICATES CORRECT ANSWER)						AVG. ANSWER TIME
	DIFF (P)	UPPER	LOWER				A	B	C	D	E	UNANSWERED	
16	0.89	95.26%	76.36%	0.19	0.18	B	2	*62	6	0	0	0	00.22
						Percentage Selected	2.86	88.57	8.57	0.00	0.00	0.0	
						Point Biserial	−0.01	0.18	−0.15	0.00	0.00		
						Disc. Index	−0.03	0.19	−0.16	0.00	0.00		
						Upper 27%	0.00	0.89	0.11	0.00	0.00		
						Lower 27%	0.03	0.70	0.27	0.00	0.00		

Q: For each type of insulin in Column A, select its peak action from Column B.

- Column A

Column B
14. Detemir

a. Long acting
15. Glargine

b. Intermediate acting
16. NPH

c. Short acting
17. Regular

What is peak action for: NPH

- A. Long acting
- *B. Intermediate acting
- C. Short acting

EXHIBIT 12.3

EXAMPLE OF A MATCHING EXERCISE (continued)

QUESTION #	CORRECT RESPONSES			DISC. INDEX	POINT BISERIAL	CORRECT ANSWER	RESPONSE FREQUENCIES (*INDICATES CORRECT ANSWER)						AVG. ANSWER TIME
	DIFF (P)	UPPER	LOWER				A	B	C	D	E	UNANSWERED	
17	0.86	88.32%	57.24%	0.31	0.28	C	2	8	*60	0	0	0	00.19
						Percentage Selected	2.86	11.43	85.71	0.00	0.00	0.0	
						Point Biserial	−0.01	−0.15	0.28	0.00	0.00		
						Disc. Index	−0.12	−0.19	0.31	0.00	0.00		
						Upper 27%	0.00	0.12	0.88	0.00	0.00		
						Lower 27%	0.12	0.31	0.57	0.00	0.00		

Q: For each type of insulin in Column A, select its peak action from Column B.

Column A

14. Detemir

15. Glargine

16. NPH

17. Regular

What is peak action for: Regular

A. Long acting

B. Intermediate acting

*C. Short acting

Column B

a. Long acting

b. Intermediate acting

c. Short acting

Source: Reprinted by permission of ExamSoft Worldwide, Inc., Dallas, TX.

EXHIBIT 12.4

EXAMPLE OF A MULTIPLE-CHOICE ITEM

ExamSoft													
QUESTION #	CORRECT RESPONSES			DISC. INDEX	POINT BISERIAL	CORRECT ANSWER	RESPONSE FREQUENCIES (*INDICATES CORRECT ANSWER)						AVG. ANSWER TIME
	DIFF (P)	UPPER	LOWER				A	B	C	D	E	UNANSWERED	
31	0.33	73.21%	89.02%	-0.16	0.12	D	9	8	30	*23	0	0	00.41
						Percentage Selected	12.86	11.43	42.86	32.86	0.00	0.00	
						Point Biserial	-0.03	0.01	0.14	0.12	0.00		
						Disc. Index	0.04	-0.18	0.30	-0.16	0.00		
						Upper 27%	0.15	0.00	0.55	0.30	0.00		
						Lower 27%	0.11	0.18	0.25	0.46	0.00		

Q: A common side effect of metformin (Glucophage) therapy is:

A. constipation.

B. hypoglycemia.

C. lactic acidosis.

*D. weight gain.


Source: Reprinted by permission of ExamSoft Worldwide, Inc., Dallas, TX.

metformin but it is a life-threatening side effect. High-scoring students may have studied this content in greater depth and therefore were attracted to this response (although this also suggests that they did not read the item carefully). The teacher should consider revising this item to change “lactic acidosis” to “metabolic acidosis” for future tests.

Item 48 (Exhibit 12.5) is a multiple-response (select all that apply) item with five response options. Only 2 of 70 students chose combination E, and a closer look at this combination reveals that including mutually exclusive alternatives might have been an unintentional clue that it was incorrect, apparent even to students who were in the low-scoring group. This combination should be revised before future use of this item. There is no obvious reason for the slightly negative discrimination index, but a large number of students selected the incorrect combination A, and some of them likely were high-scoring students. Comparing combination A with the correct answer reveals that option A excludes alternative 2, a statement about food intake. Students may have eliminated this alternative because the first sentence of the item stem focuses on exercise. This item also should be reviewed with students during the posttest discussion to determine why they chose this response instead of the correct

EXHIBIT 12.5

EXAMPLE OF A MULTIPLE-RESPONSE ITEM (SELECT ALL THAT APPLY)

ExamSoft

Questions

Rubrics

Assessments

Categories

Reports

Exam Takers

Admin

Q: A middle-aged patient newly diagnosed with type 2 diabetes wants to start an exercise program. Which of the following statements should the nurse include in this patient's teaching?

1. "Exercise increases the body's ability to metabolize glucose."

2. "If you are unable to eat due to illness, you may continue to take your antidiabetic agent with frequent glucose monitoring."

3. "If you exercise vigorously in the afternoon, be sure to eat dinner or you may have hypoglycemia in the evening or at night."

4. "Strenuous exercise is contraindicated for most patients with type 2 diabetes because of higher risk of hypoglycemic episodes."

A. 1, 3

B. 2, 4

*C. 1, 2, 3

D. 1, 2, 4

E. 1, 2, 3, 4

QUESTION #	CORRECT RESPONSES			DISC. INDEX	POINT BISERIAL	CORRECT ANSWER	RESPONSES FREQUENCIES (*INDICATES CORRECT ANSWER)						AVG. ANSWER TIME
	DIFF (P)	UPPER	LOWER				A	B	C	D	E	UNANSWERED	
48	0.41	56.23%	62.32%	−0.07	−0.08	C	23	11	*29	5	2	0	00.41
						Percentage Selected	32.86	15.71	41.43	7.14	2.86	0.00	
						Point Biserial	0.05	0.01	−0.18	0.00	−0.04		
						Disc. Index	0.08	0.02	−0.07	0.00	−0.03		
						Upper 27%	0.38	0.04	0.56	0.02	0.00		
						Lower 27%	0.30	0.02	0.63	0.02	0.03		

Source: Reprinted by permission of ExamSoft Worldwide, Inc., Dallas, TX.

one. In this case, however, the teacher should not accept combination A as a second correct response because it is not complete.

■ Conducting Posttest Discussions

Giving students feedback about test results can be an opportunity to reinforce learning, to correct misinformation, and to solicit their input for improvement of test items. But a feedback session also can be an invitation to engage in battle,

with students attacking to gain extra points and the teacher defending the honor of the test and, it often seems, the very right to give tests. Discussions with students about the test should be rational rather than opportunities for the teacher to assert power and authority. Posttest discussions can be beneficial to both teachers and students if they are planned in advance and not emotionally charged. The teacher should prepare for a posttest discussion by completing a test analysis and an item analysis and reviewing the items that were most difficult for the majority of students.

Teachers may use this information about item effectiveness as an aid to posttest discussion. The items with the lowest difficulty index (the ones answered incorrectly by the largest number of students) can be discussed at greater length, and the teacher can ask students why they selected the correct or wrong answer for such items. A discussion of the rationale for their choices may reveal common errors and misconceptions that may be corrected at that time, serve as a basis for remedial study, or contribute to a revision of those items (Ierardi, 2014).

Ierardi (2014) described a student-centered approach to posttest exam review in which a student representative volunteers to moderate the discussion of test items, with a faculty member present to clarify. Students who answered test items correctly provide insight into how they approached the items and chose the correct responses. The faculty member benefits from hearing the student discussion because it often reveals important information about items that can be used to revise them for future use.

Teachers usually assume that students choose correct responses to selection-type items because they know the content at the expected cognitive level, but there could be many other reasons for their selections. For example, a student:

- Has partial knowledge that supports choosing the correct answer
- Uses unintended cues given in the item
- Uses information from other test items
- Makes a lucky blind guess
- Intends to choose a distractor but makes a lucky clerical error in recording the correct answer

Many students who choose incorrect responses lack knowledge of the content, but other reasons for their selections may include:

- Partial knowledge that favors a distractor
- Misinformation that supports a distractor
- Unlucky blind guessing
- Intending to choose the correct answer but making a mistake in recording an incorrect response

Discussing why students chose correct or incorrect answers can reveal students' thought processes that can contribute to better teaching, improved test construction skills, and better test-taking skills.

To use time efficiently, the teacher should read the correct answers aloud quickly. If the test is hand-scored, correct answers also may be indicated by the teacher on the students' answer sheets or test booklets. If machine-scoring is used, the answer key may be projected as a scanned document from a computer or via a document camera or overhead projector. Many electronic scoring applications allow an option for marking the correct or incorrect answers directly on each student's answer sheet.

Teachers should continue to protect the security of the test during the posttest discussion by accounting for all test booklets and answer sheets and by eliminating other opportunities for cheating. Some teachers do not allow students to use pens or pencils during the feedback session to prevent answer-changing and subsequent complaints that scoring errors were made. Another approach is to distribute pens with red or green ink and permit only those pens to be used to mark answers. Teachers also should decide in advance whether to permit students to take notes during the session.

During test administration, some teachers allow students to record their answers on the test booklets, where the students also record their names, as well as on a separate answer sheet. At the completion of the exam, students submit the answer sheets and their test booklets to the teacher. When all students have finished the exam, they return to the room to check their answers using only their test booklets. The teacher might project the answers onto a screen as described previously. At the conclusion of this session, the teacher collects the test booklets again. It is important not to review and discuss individual items at this time because the test has not yet been scored and analyzed. However, the teacher may ask students to indicate problematic items and give a rationale for their answers. As discussed earlier, the teacher can use this item in conjunction with the item-analysis results to evaluate the effectiveness of test items. One disadvantage to this method of giving posttest feedback is that because the test has not yet been scored and analyzed, the teacher would not have an opportunity to thoroughly prepare for the session; feedback consists only of the correct answers, and no discussion takes place. With item effectiveness information, the teacher can identify and point out defective test items and discuss how they will be treated in scoring, rather than feel the need to defend the fairness of the items without data to support it.

Whatever the structure of the posttest discussion, the teacher should control the session so that it produces maximum benefit for all students. While discussing an item that was answered incorrectly by a majority of students, the teacher should maintain a calm, matter-of-fact, nondefensive attitude. The teacher should avoid arguing with students about individual items and engaging in emotionally charged discussion; instead, the teacher should either invite written comments as described previously or schedule individual appointments to discuss the items in question.

Students who need additional help also may be encouraged to make appointments with the teacher for individual review sessions.

Eliminating Items or Adding Points

Teachers often debate the merits of adjusting test scores by eliminating items or adding points to compensate for real or perceived deficiencies in test construction or performance. For example, during a posttest discussion, students may argue that if they all answered an item incorrectly, the item should be omitted or all students should be awarded an extra point to compensate for the “bad item.” It is interesting to note that students seldom propose subtracting a point from their scores if they all answer an item correctly. In any case, how should the teacher respond to such requests? In this discussion, a distinction is made between test items that are technically flawed and those that do not function as intended.

If test items are properly constructed, critiqued, and proofread, it is unlikely that serious flaws will appear on the test. However, errors that do appear may have varying effects on students’ scores. For example, if the correct answer to a multiple-choice item is inadvertently omitted from the test, no student will be able to answer the item correctly. In this case, the item simply should not be scored. That is, if the error is discovered during or after test administration and before the test is scored, the item is omitted from the answer key; a test that was intended to be worth 73 points then is worth 72 points. If the error is discovered after the tests are scored, they can be rescored. Students often worry about the effect of this change on their scores and may argue that they should be awarded an extra point in this case. The possible effects of both adjustments on a hypothetical score are shown in Table 12.4.

It is obvious that omitting the flawed item and adding a point to the raw score produce nearly identical results. Although students might view adding a point to their scores as more satisfying, it makes little sense to award a point for an item that was not answered correctly. The “extra” point in fact does not represent knowledge of any

TABLE 12.4 Effects of Test Score Adjustments

	TOTAL POSSIBLE POINTS	RAW SCORE	PERCENTAGE CORRECT
Original test	73	62	84.9
Flawed item not scored	72	62	86.1
Point added to raw score	73	63	86.3

content area or achievement of an objective, and therefore it does not contribute to a valid interpretation of the test scores. Teachers should inform students matter-of-factly that an item was eliminated from the test and reassure them that their relative standing with regard to performance on the test has not changed.

If the technical flaw consists of a misspelled word in a true–false item that does not change the meaning of the statement, no adjustment should be made. The teacher should avoid lengthy debate about item semantics if it is clear that such errors are unlikely to have affected the students’ scores. Feedback from students can be used to revise items for later use and sometimes make changes in teaching that concept or skill.

As previously discussed, teachers should resist the temptation to eliminate items from the test solely on the basis of low difficulty and discrimination indices. Omission of items may affect the validity of the scores from the test, particularly if several items related to one content area or objective are eliminated, resulting in inadequate sampling of that content (Miller et al., 2013). This is particularly true for quizzes that contain a small number of items.

Because identified flaws in test construction do contribute to measurement error, the teacher should consider taking them into account when using the test scores to make grading decisions and set cutoff scores. That is, the teacher should not fix cutoff scores for assigning grades until after all tests have been given and analyzed. The proposed grading scale can then be adjusted if necessary to compensate for deficiencies in test construction. It should be made clear to students that any changes in the grading scale because of flaws in test construction would not adversely affect their grades.

■ Developing a Test-Item Bank

Because considerable effort goes into developing, administering, and analyzing test items, teachers should develop a system for maintaining and expanding a pool or bank of items from which to select items for future tests. Teachers can maintain databases of test items on their computers with backups on storage devices. When teachers store test-item databases electronically, the files must be password-protected and test security maintained. When developing test banks, the teacher can record the following data with each test item: (a) the correct response for objective-type items and a brief scoring key for completion or essay items; (b) the course, unit, content area, or objective for which it was designed; and (c) the item-analysis results for a specified period of time. Exhibit 12.6 offers one such example.

Commercially produced software applications can be used in a similar way to develop and store a database of test items. Each test item is a record in the database. The test items can then be sorted according to the fields in which the data are entered;

EXHIBIT 12.6**SAMPLE INFORMATION TO INCLUDE WITH ITEMS IN THE TEST BANK**

Content Area: Physical Assessment

Unit 5

Objective 3

1. What is the most likely explanation for breast asymmetry in an adolescent girl?
 - A. Blocked mammary duct in the larger breast
 - B. Endocrine disorder
 - C. Mastitis in the larger breast
 - D. Normal variation in growth¹

TEST DATE	DIFF. INDEX	DISCRIM. INDEX
10/22	.72	.25
2/20	.56	.33
8/23	.60	.40

Note: ¹ = correct answer

Diff. Index, difficulty index; Discrim. Index, discrimination index.

for example, the teacher could retrieve all items that are classified as Objective 3, with a moderate difficulty index.

Many publishers also offer test-item banks that relate to the content contained in their textbooks. However, faculty members need to be cautious about using these items for their own examinations. The purpose of the test, relevant characteristics of the students to be tested, and the balance and emphasis of content as reflected in the teacher's test blueprint are the most important criteria for selecting test items. Although some teachers would consider these item banks to be a shortcut to test development, items should be evaluated carefully before they are used. There is no guarantee that the quality of test items in a published item bank is superior to that of test items that a skilled teacher can construct. Many of the items may be of questionable quality. Often, a teacher can improve the quality of commercial test-bank items that are congruent with his or her test blueprint by modifying an item stem, substituting better answer options, eliminating technical flaws, or changing the item format.

In addition, published test-item banks seldom contain item-analysis information such as difficulty and discrimination indices. However, the teacher can calculate this information for each item used or modified from a published item bank, and can develop and maintain an item file.

■ Summary

To accurately interpret test scores, the teacher needs to analyze the performance of the test as a whole as well as the individual test items. Information about how the test performed helps teachers to give feedback to students about test results and to improve test items for future use.

Scoring a test results in a collection of numbers known as *raw scores*. To make raw scores understandable, they can be arranged in frequency distributions or displayed graphically as histograms or frequency polygons. Score distribution characteristics such as symmetry, skewness, modality, and kurtosis can assist the teacher in understanding how the test performed as a measurement tool as well as to interpret any one score in the distribution.

Measures of central tendency and variability also aid in interpreting individual scores. Measures of central tendency include the mode, median, and mean; each measure has advantages and disadvantages for use. In a normal distribution, these three measures will coincide. Most score distributions from teacher-made tests do not meet the assumptions of a normal curve. The shape of the distribution can determine the most appropriate index of central tendency to use. Variability in a distribution can be described roughly as the range of scores or, more precisely, as the standard deviation.

Teachers can make criterion-referenced or norm-referenced interpretations of individual student scores. Norm-referenced interpretations of any individual score should take into account the characteristics of the score distribution, some index of central tendency, and some index of variability. The teacher thus can use the mean and standard deviation to make judgments about how an individual student's score compares with those of others.

A percentage-correct score is calculated by dividing the raw score by the total possible score; thus, it compares the student's score to a preset standard or criterion. A percentage-correct score is not an objective indication of how much a student really knows about a subject because it is affected by the difficulty of the test items. The percentage-correct score should not be confused with percentile rank, which describes the student's relative standing within a group and therefore is a norm-referenced interpretation. The percentile rank of a given raw score is the percentage of scores in the distribution that occurs at or below that score. The results of standardized tests usually are reported as percentile ranks or other norm-referenced scores. Teachers should be cautious when interpreting standardized test results so that comparisons with the appropriate norm group are made. Standardized test scores should not be used to determine grades, and results should be interpreted with the understanding that only large differences in scores indicate real differences in achievement levels.

Item analysis typically is performed by the use of a computer program, either as part of a test scoring application or computer testing software. The difficulty index (P), ranging from 0 to 1.00, indicates the percentage of students who answered the item correctly. Items with P -values of .20 and below are considered to be difficult, and those with P -values of .80 and above are considered to be easy. However, interpretation of the difficulty index should take into account the quality of the instruction and the abilities of the students in the group. The discrimination index (D), ranging from -1.00 to $+1.00$, is an indication of the extent to which high-scoring students answered the item correctly more often than low-scoring students did. In general, the higher the positive value, the better the test item; desirable discrimination indexes should be at least $+.20$. An item's power to discriminate is highly related to its difficulty index. An item that is answered correctly by all students has a difficulty index of 1.00; the discrimination index for this item is 0, because there is no difference in performance on that item between high scorers and low scorers.

Flaws in test construction may have varying effects on students' scores and therefore should be handled differently. If the correct answer to a multiple-choice item is inadvertently omitted from the test, no student will be able to answer the item correctly. In this case, the item simply should not be scored. If a flaw consists of a misspelled word that does not change the meaning of the item, no adjustment should be made.

Teachers should develop a system for maintaining a pool or bank of items from which to select items for future tests. Item banks frequently are a feature of computer testing programs, or they can be developed by the faculty and stored electronically. Use of published test-item banks should be based on the teacher's evaluation of the quality of the items as well as on the purpose for testing, relevant characteristics of the students, and the desired emphasis and balance of content as reflected in the teacher's test blueprint. Items selected from a published item bank often must be modified to be technically sound and relevant to how the content area was taught and the characteristics of the students to be tested.

■ References

- Brookhart, S. M., & Nitko, A. J. (2019). *Educational assessment of students* (8th ed.). New York, NY: Pearson Education.
- Ermie, E. (n.d.). *Exam quality through the use of psychometric analysis [White Paper]*. Dallas, TX: ExamSoft Worldwide. Retrieved from <https://examsoft.com/resources/white-papers/exam-quality-use-psychometric-analysis>
- Ierardi, J. A. (2014). Taking the "sting" out of examination reviews: A student-centered approach. *Journal of Nursing Education*, 53, 428.
- Miller, M. D., Linn, R. L., & Gronlund, N. E. (2013). *Measurement and assessment in teaching* (11th ed.). Upper Saddle River, NJ: Prentice Hall.
- Waltz, C. F., Strickland, O. L., & Lenz, E. R. (2010). *Measurement in nursing and health research* (4th ed.). New York, NY: Springer Publishing Company.

IV

CLINICAL EVALUATION

CLINICAL EVALUATION PROCESS

Nursing as a practice discipline requires development of higher level cognitive skills, values, and psychomotor and technical skills for care of patients across settings. Acquisition of knowledge alone is not sufficient; professional education includes a practice dimension in which students develop competencies for care of patients and learn to think like professionals. Through clinical evaluation, the teacher arrives at judgments about the students' competencies—their performance in practice. This chapter describes the process of clinical evaluation in nursing; in Chapter 14, Clinical Evaluation Methods, specific clinical evaluation methods are presented.

■ Outcomes of Clinical Practice

There are many outcomes that students can achieve through their clinical practice experiences. In clinical courses, students acquire knowledge and learn about concepts and evidence to guide their patient care. They have an opportunity to transfer learning from readings, face-to-face classes and discussions, online classes, simulations, and other experiences to care of patients.

Clinical experiences provide an opportunity for students to use research findings and other evidence to make decisions about interventions and other aspects of patient care. In the practice setting, they learn the process of evidence-based nursing and how to translate evidence to patient care. They also need to acquire knowledge, skills, and attitudes for improving the quality of healthcare (Altmiller & Armstrong, 2017; Dolansky & Moore, 2013; Drenkard, 2015; Interprofessional Education Collaborative Expert Panel, 2016; Johnson, Drenkard, Emard, & McGuinn, 2015; Phillips, Stalter, Dolansky, & Lopez, 2016; Quality and Safety Education for Nurses [QSEN], 2019). In practice, students deal with ambiguous patient situations and unique cases that do not fit the textbook description; this requires students to think critically about what to do. For this reason, clinical practice, whether in the patient care setting or simulation laboratory, is important for developing higher level cognitive skills and for learning to arrive at clinical judgments based on available

information (Manetti, 2018, 2019; Victor, Ruppert, & Ballasy, 2017). Clinical experiences present problems and situations that may not lend themselves to resolution through the application of research findings and theories learned in class and through one's readings. When faced with problems in clinical practice that are not clear-cut or easily solved, students have an opportunity to develop their thinking and clinical judgment skills.

Through practice experiences with patients and in learning and simulation laboratories, students develop their psychomotor skills, learn how to use technology, and gain necessary skills for implementing nursing and other interventions. This practice is essential for initial learning, to refine competencies, and to maintain them over a period of time. Through practice students also learn the realities of caring for patients in varied healthcare environments. As healthcare systems and patients rely increasingly on information technology, students must acquire informatics competencies and be able to use new technologies, such as telehealth and digital health tools, as they are introduced into practice.

Having technical skills, though, is only one aspect of professional practice. In caring for patients and working with nurses and other healthcare providers, students gain an understanding of how professionals approach their patients' problems, how they interact with each other, and what behaviors are important in carrying out their roles and working as a team in the practice setting. Learning to collaborate with other healthcare professionals and function effectively on nursing and interprofessional teams are critical to providing quality and safe care (Horsley et al., 2016; Interprofessional Education Collaborative Expert Panel, 2016; Sullivan, Kiovisky, Mason, Hill, & Dukes, 2015). Clinical learning activities provide an opportunity for students to develop their individual and team communication skills and learn how to collaborate with others.

Practice as a professional is contingent not only on having knowledge to guide decisions but also on having a value system that recognizes the worth, dignity, and rights of patients and others in the healthcare system. As part of this value system, students need to develop cultural competence and gain the knowledge and attitudes essential to provide multicultural healthcare. As society becomes more diverse, it is critical for nursing students to become culturally competent (Blanchet Garneau, 2016; de Castro, Dyba, Cortez, & Pe Benito, 2019; Dyches, Haynes-Ferere, & Haynes, 2019; Flood & Commendador, 2016). Much of this learning can occur in clinical practice as students care for culturally diverse patients and communities and through simulations in which they can explore cultural differences. Clinical experiences help students develop competencies in patient-centered care: respecting patients' preferences, values, and needs; recognizing patients as partners in care; providing compassionate care; continuously coordinating care; and advocating for patients (Bentley,

Engelhardt, & Watzak, 2014; Horsley et al., 2016). These core competencies, needed by all healthcare professionals, are developed in clinical practice.

Another outcome of clinical practice is developing knowledge, skills, and values to continuously improve the quality and safety of healthcare (Altmiller & Armstrong, 2017; Drenkard, 2015; Johnson et al., 2015; Phillips et al., 2016; QSEN, 2019). Nursing students need to learn quality-improvement methods and have experience with them as part of their clinical practice. They also need to understand their role in creating a safe healthcare system for patients and a safety culture in every clinical setting, learn about healthcare errors and how to prevent them, and value the importance of error reporting. These are competencies that can be developed in simulation and clinical practice.

Some clinical courses focus on management and leadership outcomes. For those courses, clinical practice provides learning opportunities for students to manage groups of patients, provide leadership in the healthcare setting, and learn how to delegate, among other competencies.

In clinical practice, students learn to accept responsibility for their actions and decisions about patients. They also should be willing to accept errors in judgment and learn from them. These are important outcomes of clinical practice in any nursing and health professions program.

Another outcome of clinical practice is learning to learn. Professionals in any field are learners throughout the duration of their careers. Continually expanding knowledge, developments in healthcare, and new technology alone create the need for lifelong learners in nursing. In clinical practice, students are faced with situations of which they are unsure; they are challenged to raise questions about patient care and seek further learning. In nursing courses as students are faced with gaps in their learning, they should be guided in the self-assessment process, directed to resources for learning, and supported by the teacher. All too often students are hesitant to express their learning needs to their teachers for fear of the effect it will have on their grade or on the teacher's impression of the student's competence in clinical practice.

These outcomes of clinical practice are listed in Exhibit 13.1. Integrated in this list are the core competencies needed by all healthcare professionals: patient-centered care, teamwork and collaboration, evidence-based practice, quality improvement, safety, and informatics (Dolansky & Moore, 2013; QSEN, 2019). The outcomes provide a framework for faculty members to use in planning their clinical courses and deciding how to assess student performance. Not all outcomes are applicable to every nursing course; for instance, some courses may not call for the acquisition of technological or delegation skills, but overall, most courses will move students toward achievement of these outcomes as they progress through the nursing program.

EXHIBIT 13.1**OUTCOMES OF CLINICAL PRACTICE IN NURSING PROGRAMS**

- ☐ Acquire concepts, theories, and other knowledge for clinical practice.
- ☐ Use research and other evidence in clinical practice.
- ☐ Develop higher level thinking and clinical judgment skills.
- ☐ Develop psychomotor and technical skills, competence in performing other types of interventions, and informatics competencies.
- ☐ Communicate effectively with patients, families, communities, and others in the healthcare system.
- ☐ Collaborate with and lead interprofessional teams.
- ☐ Develop values and knowledge essential for providing patient-centered care to a culturally and ethnically diverse patient population.
- ☐ Develop knowledge, skills, and values essential for continuously improving the quality and safety of healthcare.
- ☐ Demonstrate leadership skills and behaviors of a professional.
- ☐ Accept responsibility for actions and decisions.
- ☐ Accept the need for continued learning and self-development.

■ Concept of Clinical Evaluation

Clinical evaluation is a process by which judgments are made about learners' competencies in practice. This practice may involve care of patients, families, and communities; other types of experiences in the clinical setting; simulated experiences; and performance of varied skills. Most often, clinical evaluation involves observing performance and arriving at judgments about the student's competence. Judgments influence the data collected, that is, the specific types of observations made to evaluate the student's performance, and the inferences and conclusions drawn from the data about the quality of that performance. Teachers may collect different data to evaluate the same outcomes, and when presented with a series of observations about a student's performance in clinical practice, there may be little consistency in their judgments about how well that student performed.

Clinical evaluation is not an objective process; it is subjective—involving judgments by the teacher and others involved in the process. As discussed in Chapter 1, Assessment and the Educational Process, the teacher's values influence evaluation. This is most apparent in clinical evaluation, where our values influence the observations we make of students and the judgments we make about the quality of their performance. Thus, it is important for teachers to be aware of their own values that might bias their judgments of students.

This is not to suggest that clinical evaluation can be value-free; the teacher's observations of performance and conclusions always will be influenced by her or his values. The key is to develop an awareness of these values so as to avoid their influencing clinical evaluation to a point of unfairness to the student. For example, if the teacher prefers students who initiate discussions and participate actively in conferences, this value should not influence judgments about students' competencies in other areas. The teacher needs to be aware of this preference to avoid an unfair evaluation of other dimensions of the students' clinical performance. Or, if the teacher is used to the fast pace of most acute care settings, when working with beginning students or someone who "moves slowly," the teacher should be cautious not to let this prior experience influence expectations of performance. Clinical educators should examine their own values, attitudes, and beliefs so that they are aware of them as they teach and assess students' performance in practice settings.

Clinical Evaluation Versus Grading

Clinical evaluation is not the same as grading. In evaluation, the teacher makes observations of performance and collects other types of data, then compares this information to a set of standards to arrive at a judgment. From this assessment, a quantitative symbol or grade may be applied to reflect the evaluation data and judgments made about performance. The clinical grade, such as pass-fail or A through F, is the symbol used to represent the evaluation. Clinical performance may be evaluated and not graded, such as with formative evaluation or feedback to the learner, or it may be graded. Grades, however, should not be assigned without sufficient data about clinical performance.

Norm- and Criterion-Referenced Clinical Evaluation

Clinical evaluation may be either norm referenced or criterion referenced, as described in Chapter 1, Assessment and the Educational Process. In norm-referenced evaluation, the student's clinical performance is compared with that of other students, indicating that the performance is better than, worse than, or equivalent to that of others in the comparison group or that the student has more or less knowledge, skill, or ability than the other students. Rating students' clinical competencies in relation to others in the clinical group—for example, indicating that the student was "average"—is a norm-referenced interpretation.

In contrast, criterion-referenced clinical evaluation involves comparing the student's clinical performance with predetermined criteria, not to the performance of other students in the group. In this type of clinical evaluation, the criteria are known in advance and used as the basis for evaluation. Indicating that the student has met

the clinical outcomes or achieved the clinical competencies, regardless of how other students performed, represents a criterion-referenced interpretation.

Formative and Summative Clinical Evaluation

Clinical evaluation may be formative or summative. Formative evaluation in clinical practice provides feedback to learners about their progress in meeting the outcomes of the clinical course or in developing the clinical competencies. The purposes of formative evaluation are to enable students to develop further their clinical knowledge, skills, and values; indicate areas in which learning and practice are needed; and provide a basis for suggesting additional instruction to improve performance. With this type of evaluation, after identifying the learning needs, instruction is provided to move students forward in their learning. Formative evaluation, therefore, is diagnostic; it should not be graded (Brookhart & Nitko, 2019). For example, the clinical teacher or preceptor might observe a student perform wound care and give feedback on changes to make with the technique. The goal of this assessment is to improve subsequent performance, not to grade how well the student carried out the procedure.

Summative clinical evaluation, however, is designed for determining clinical grades because it summarizes competencies the student has developed in clinical practice. Summative evaluation is done at the end of a period of time, for example, at midterm or at the end of the clinical practicum, to assess the extent to which learners have achieved the clinical outcomes or competencies. Summative evaluation is not diagnostic; it summarizes the performance of students at a particular point in time. For much of clinical practice in a nursing education program, summative evaluation comes too late for students to have an opportunity to improve performance. Any protocol for clinical evaluation should include extensive formative evaluation and periodic summative evaluation. Formative evaluation is essential to provide feedback to improve performance while practice experiences are still available.

■ Fairness in Clinical Evaluation

Considering that clinical evaluation is not objective, the goal is to establish a *fair* evaluation system. Fairness requires that:

1. The clinical teacher identify his or her own values, attitudes, beliefs, and biases that may influence the evaluation process.
2. Clinical evaluation is based on predetermined outcomes or competencies.
3. The teacher develops a supportive clinical learning environment.

Identify One's Own Values

Teachers need to be aware of their personal values, attitudes, beliefs, and biases, which may influence the evaluation process. These can affect both the data collected about students and the judgments made about performance. In addition, students have their own set of values and attitudes that influence their self-evaluations of performance and their responses to the teacher's evaluations and feedback. Students' acceptance of the teacher's guidance in clinical practice and information provided to them for improving performance is affected by their past experiences in clinical courses with other faculty. Students may have had problems in prior clinical courses, receiving only negative feedback and limited support from the teacher, staff members, and others. In situations in which student responses inhibit learning, the teacher may need to intervene to guide students to be more self-aware concerning the students' own values and the effect they are having on learning.

Base Clinical Evaluation on Predetermined Outcomes or Competencies

Clinical evaluation should be based on preset outcomes or clinical competencies that are then used to guide the evaluation process. Without these, neither the teacher nor the student has any basis for evaluating clinical performance. What are the outcomes of the clinical course to be met or what competencies should the student achieve in this clinical practicum? These outcomes or competencies provide a framework for educators to use in observing performance and for arriving at judgments about achievement in clinical practice. For example, if the competencies relate to developing communication skills, then the learning activities, whether in the patient care setting or as part of a simulation, should assist students in learning how to communicate. The teacher's observations and subsequent assessment should focus on communication behaviors, not on other competencies unrelated to the learning activities.

Develop a Supportive Learning Environment

It is up to the teacher to develop a supportive learning environment in which students view the teacher as someone who will facilitate their learning and development of clinical competencies. Students need to be comfortable asking faculty and staff members questions and seeking their guidance rather than avoiding them in the clinical setting. A supportive environment is critical to effective assessment because students need to recognize that the teacher's feedback is intended to help them improve performance. Developing a "climate" for learning is also important because clinical practice is stressful for students

(Bagcivan, Cinar, Tosun, & Korkmaz, 2015; Bhurtun, Azimirad, Saaranen, & Turunen, 2019; Blomberg et al., 2014; Suresh, Matthews, & Coyne, 2013; Zieber & Williams, 2015). Many factors influence the development of this learning climate. The clinical setting needs to provide experiences that foster student learning and development. Staff members need to be supportive of students; work collaboratively with each other, students, and the clinical teacher; and communicate effectively, both individually and as a team. Most of all, trust and respect must exist between the teacher and the students.

■ Student Stress in Clinical Practice

There have been a number of studies in nursing education on student stress in the clinical setting. Some of the stresses students have identified are:

- The fear of making a mistake that would harm the patient
- Having insufficient knowledge and skills for patient care
- Changing patient conditions and uncertainty about how to respond
- Being unfamiliar with the staff, policies, and other aspects of the clinical setting
- Caring for difficult patients
- Having the teacher observe and evaluate clinical performance
- Interacting with the patient, the family, nursing staff, and other healthcare providers

Learning in the clinical setting is a *public experience*. Students cannot hide their lack of understanding or skills as they might in class or in an online discussion. In clinical practice, the possibility exists for many people to observe the student's performance—the teacher, patient, family members, peers, nursing staff, and other healthcare providers. Being observed and evaluated by others is stressful for students in any healthcare field.

The potential stress that students might experience in clinical practice reinforces the need for faculty members to be aware of the learning environment they set when working with students in a clinical course. The student is a learner, not a nurse, although some educators, preceptors, and other providers expect students to perform at an expert level without giving them sufficient time to practice and refine their performance (Oermann, Shellenbarger, & Gaberson, 2018). Simulated experiences may be effective in reducing some of the anxieties students experience by allowing them to practice their skills, both cognitive and psychomotor, prior to care of patients.

■ Feedback in Clinical Evaluation

For clinical evaluation to be effective, the teacher should provide continuous feedback to students about their performance and how they can improve it. *Feedback* is the communication of information to students, based on the teacher's assessment, that enables students to reflect on their performance, identify continued learning needs, and decide how to meet them (Bonnell, 2008). Feedback may be verbal, by describing observations of performance and explaining what to do differently, or visual, by demonstrating correct performance. Feedback should be specific and accompanied by further instruction from the teacher or by working with students to identify appropriate learning activities. The ultimate goal is for students to progress to a point at which they can judge their own performance, identify resources for their learning, and use those resources to further develop competencies. Bonnell (2008) emphasized that for feedback to be useful, students need to reflect on the information communicated to them and take an active role in incorporating that feedback in their own learning (p. 290).

Students must have an underlying knowledge base and beginning skills to judge their own performance. Brookhart and Nitko (2019) suggested that feedback on performance also identifies the possible causes or reasons why the student has not mastered the learning outcomes. Sometimes, the reason is that the student does not have the prerequisite knowledge and skills for developing the new competencies. As such it is critical for clinical teachers and preceptors to begin their interactions with students by assessing whether students have learned the necessary concepts and skills and, if not, to start there.

Principles of Providing Feedback as Part of Clinical Evaluation

There are five principles for providing feedback to students as part of the clinical evaluation process. First, the feedback should be precise and specific. General information about performance, such as “You need to work on your assessment” or “You need more practice in the simulation center,” does not indicate which behaviors need improvement or how to develop them. Instead of using general statements, the teacher should indicate what specific areas of knowledge are lacking, where there are problems in thinking and clinical judgments, and what particular competencies need more development. Rather than saying to a student, “You need to work on your assessment,” the feedback would be more effective if the teacher identified the specific areas of data collection omitted and the physical examination techniques that need improvement. Specific feedback is more valuable to learners than a general description of their behavior.

Second, for procedures, use of technologies, and psychomotor skills, the teacher should provide both verbal and visual feedback to students. This means that the

teacher should explain first where the errors were made in performance and then demonstrate the correct procedure or skill. Research suggests that physically guiding learners in how to perform the procedure or skill improves their accuracy (Soderstrom & Bjork, 2015). This should be followed by the student practicing the skill with the teacher guiding performance. By allowing immediate practice, with the teacher available to correct problems, students can more easily *use* the feedback to further develop their skills.

Third, feedback about performance should be given to students at the time of learning or immediately following it. Giving prompt feedback is one of the seven core principles for effective teaching in undergraduate programs (Chickering & Gamson, 1987). Providing prompt and rich feedback is equally important when teaching graduate students, nurses, and other learners regardless of their educational level. The longer the period of time between performance and feedback from the teacher, the less effective the feedback (Oermann et al., 2018). As time passes, neither student nor teacher may remember specific areas of clinical practice to be improved. This principle holds true whether the performance relates to clinical judgment or other cognitive skills, a procedure or technical skill, or an attitude or value expressed by the student, among other areas. Whether working with a group of students in a clinical setting, communicating with preceptors about students, or teaching an online course, the teacher needs to develop a strategy for giving focused and prompt feedback to students and following up with further discussion as needed. Recording short notes for later discussion with individual students may help the teacher remember important points about performance.

Fourth, students need different amounts of feedback and positive reinforcement. In beginning practice and with clinical situations that are new to learners, most students will need frequent and extensive feedback. As students progress through the program and become more competent, they should be able to assess their own performance and identify personal learning needs. Some students will require more feedback and direction from the teacher than others. As with many aspects of education, one approach does not fit all students. Feedback should always be given to students in a private area.

One final principle is that feedback should be diagnostic. This means that after identifying areas in which further learning is needed, the teacher's responsibility is to guide students so that they can improve their performance. Altmiller (2016) emphasized the importance of being attentive to how the feedback is delivered: the teacher should identify the feedback as an opportunity for student learning and include options for improvement. The process is cyclical—the teacher observes and assesses performance, gives students feedback about that performance, and then guides their learning and practice so they can become more competent.

Gigante, Dell, and Sharkey (2011) proposed a five-step process for giving feedback to students:

1. Identify the expectations for the student. Students need to know what is expected of them in the clinical practicum.
2. Set the stage for the student to receive feedback from the teacher and others involved in the learning situation. The authors recommend beginning with this phrase, “I am giving you feedback” because then students realize that the information is to help them improve performance.
3. Begin the interaction by asking students to assess their own performance, which encourages reflection and learning.
4. Describe how the student is performing based on specific observations of behaviors, which should be shared. It is important to provide concrete examples of performance and describe specifically how the learner can improve.
5. Ask for input from the learner. In some cases, such as when there are concerns about not achieving at a satisfactory level in the course, a written plan for improvement should be developed with consequences outlined.

■ Clinical Outcomes and Competencies

There are different ways of specifying the outcomes to be achieved in clinical practice, which in turn provide the basis for clinical evaluation. These may be stated in the form of outcomes to be met or as competencies to be demonstrated in clinical practice. Regardless of how these are stated, they represent *what* is evaluated in clinical practice.

The outcomes of clinical practice offered in Exhibit 13.1 can be used for developing specific outcomes or competencies for a clinical course. Not all clinical courses will have outcomes in each of these areas, and in some courses, there may be other types of competencies unique to practice in that clinical specialty. Some faculty members identify common outcomes or competencies that are used for each clinical course in the program and then level those to demonstrate their progressive development through the nursing program (Billings & Halstead, 2016). For example, with this model, each course would have an outcome on communication. In a beginning clinical course, the outcome might be, “Identifies verbal and nonverbal techniques for communicating with patients.” In a later course in the curriculum, the communication outcome might focus on the family and working with caregivers, for example, “Develops interpersonal relationships with families and caregivers.” Then in the community health course the outcome might be, “Collaborates with other providers, interdisciplinary groups, and community organizations.”

As another approach, some faculty members state the outcomes broadly and then indicate specific behaviors students should demonstrate to meet those outcomes in a particular course. For example, the outcome on communication might be stated as “Communicates effectively with patients and on intra- and interprofessional teams.” Examples of behaviors that indicate achievement of this outcome in a course on care of children include, “Uses appropriate verbal and nonverbal communication based on the child’s age, developmental status, and health condition” and “Interacts effectively with parents, caregivers, and the interprofessional team.” Generally, the outcomes or competencies are then used for developing the clinical evaluation tool or rating form, which is discussed in Chapter 14, Clinical Evaluation Methods.

Regardless of how the outcomes are stated for a clinical course, they need to be specific enough to guide the evaluation of students in clinical practice. An outcome such as “Use the nursing process in care of children” is too broad to guide evaluation. More specific outcomes such as “Carries out a systematic assessment of children reflecting their developmental stage,” “Evaluates the impact of health problems on the family,” and “Identifies resources for managing the child’s care at home” make clear to students what is expected of them in clinical practice.

Competencies are the abilities to be demonstrated by the learner in clinical practice. Competencies are the knowledge, skills, and attitudes that students need to develop; they provide the foundation for evaluation (Sullivan, 2016). For nurses in practice, these competencies reflect the expected level of performance for caring for patients in the healthcare setting. Competencies for nurses are assessed on hire and on an ongoing basis, usually annually, validating that nurses are competent to practice (Levine & Johnson, 2014). Caution should be exercised in developing clinical outcomes and competencies to avoid having too many for evaluation, considering the number of learners for whom the teacher is responsible, types of clinical learning opportunities available, and time allotted for clinical learning activities. In preparing outcomes or competencies for a clinical course, teachers should keep in mind that they need to collect sufficient data about students’ performance of each outcome or competency specified for that course. Too many outcomes make it nearly impossible to collect enough data on the performance of all of the students in the clinical setting whether they are in a small group with a faculty member on site or are working one-to-one with a clinician. Regardless of how the evaluation system is developed, the clinical outcomes or competencies need to be realistic and useful for guiding the evaluation.

■ Summary

Through clinical evaluation, the teacher arrives at judgments about students’ performance in clinical practice. The teacher’s observations of performance should focus on the outcomes to be met or competencies to be developed in the clinical course. These provide the framework for learning in clinical practice and the basis for evaluating performance.

Although a framework such as this is essential in clinical evaluation, teachers also need to examine their own beliefs about the evaluation process and the purposes it serves in nursing. Clarifying one's own values, beliefs, attitudes, and biases that may affect evaluation is an important first step. Recognizing the inherent stress of clinical practice for many students and developing a supportive learning environment are also important. Other concepts of evaluation, presented in Chapter 1, Assessment and the Educational Process, apply to clinical evaluation. Specific methods for clinical evaluation are described in Chapter 14, Clinical Evaluation Methods.

■ References

- Altmiller, G. (2016). Strategies for providing constructive feedback to students. *Nurse Educator*, 41, 118–119. doi:10.1097/nne.0000000000000227
- Altmiller, G., & Armstrong, G. (2017). National Quality and Safety Education for Nurses faculty survey results. *Nurse Educator*, 42, S3–S7. doi:10.1097/nne.0000000000000408
- Bagcivan, G., Cinar, F. I., Tosun, N., & Korkmaz, R. (2015). Determination of nursing students' expectations for faculty members and the perceived stressors during their education. *Contemporary Nurse*, 50(1), 58–71. doi:10.1080/10376178.2015.1010259
- Bentley, R., Engelhardt, J. A., & Watzak, B. (2014). Collaborating to implement interprofessional educational competencies through an international immersion experience. *Nurse Educator*, 39, 77–84. doi:10.1097/NNE.0000000000000022
- Bhurtun, H. D., Azimirad, M., Saaranen, T., & Turunen, H. (2019). Stress and coping among nursing students during clinical training: An integrative review. *Journal of Nursing Education*, 58, 266–272. doi:10.3928/01484834-20190422-04
- Billings, D. M., & Halstead, J. A. (2016). *Teaching in nursing: A guide for faculty* (5th ed.). St. Louis, MO: Elsevier.
- Blanchet Garneau, A. (2016). Critical reflection in cultural competence development: A framework for undergraduate nursing education. *Journal of Nursing Education*, 55, 125–132. doi:10.3928/01484834-20160216-02
- Blomberg, K., Bisholt, B., Kullen Engstrom, A., Ohlsson, U., Sundler Johansson, A., & Gustafsson, M. (2014). Swedish nursing students' experience of stress during clinical practice in relation to clinical setting characteristics and the organisation of the clinical education. *Journal of Clinical Nursing*, 23, 2264–2271. doi:10.1111/jocn.12506
- Bonnel, W. (2008). Improving feedback to students in online courses. *Nursing Education Perspectives*, 29, 290–294.
- Brookhart, S. M., & Nitko, A. J. (2019). *Educational assessment of students* (8th ed.). New York, NY: Pearson Education.
- Chickering, A. W., & Gamson, Z. F. (1987). Seven principles for good practice in undergraduate education. *AAHE Bulletin*, 39(7), 3–7.
- de Castro, A. B., Dyba, N., Cortez, E. D., & Pe Benito, G. G. (2019). Collaborative online international learning to prepare students for multicultural work environments. *Nurse Educator*, 44, E1–E5. doi:10.1097/nne.0000000000000609
- Dolansky, M. A., & Moore, S. M. (2013). Quality and Safety Education for Nurses (QSEN): The key is systems thinking. *Online Journal of Issues in Nursing*, 18(3), Manuscript 1. doi:10.3912/OJIN.Vol18No03Man01

- Drenkard, K. N. (2015). The power of alignment: Educating nurses in quality and safety. *Nursing Administration Quarterly*, 39, 272–277. doi:10.1097/NAQ.0000000000000112
- Dyches, C., Haynes-Ferere, A., & Haynes, T. (2019). Fostering cultural competence in nursing students through international service immersion experiences. *Journal of Christian Nursing*, 36, E29–E35. doi:10.1097/cnj.0000000000000602
- Flood, J. L., & Commendador, K. A. (2016). Undergraduate nursing students and cross-cultural care: A program evaluation. *Nurse Education Today*, 36, 190–194. doi:10.1016/j.nedt.2015.10.003
- Gigante, J., Dell, M., & Sharkey, A. (2011). Getting beyond “good job”: How to give effective feedback. *Pediatrics*, 127, 205–207.
- Horsley, T. L., Reed, T., Muccino, K., Quinones, D., Siddall, V. J., & McCarthy, J. (2016). Developing a foundation for interprofessional education within nursing and medical curricula. *Nurse Educator*, 41, 234–238. doi:10.1097/NNE.0000000000000255
- Interprofessional Education Collaborative Expert Panel. (2016). *Core competencies for interprofessional collaborative practice: 2016 update*. Washington, DC: Interprofessional Education Collaborative.
- Johnson, J., Drenkard, K., Emard, E., & McGuinn, K. (2015). Leveraging Quality and Safety Education for Nurses to enhance graduate-level nursing education and practice. *Nurse Educator*, 40, 313–317. doi:10.1097/NNE.0000000000000177
- Levine, J., & Johnson, J. (2014). An organizational competency validation strategy for registered nurses. *Journal for Nurses in Professional Development*, 30, 58–65. doi:10.1097/NND.0000000000000041
- Manetti, W. (2018). Evaluating the clinical judgment of prelicensure nursing students in the clinical setting. *Nurse Educator*, 43, 272–276. doi:10.1097/nne.0000000000000489
- Manetti, W. (2019). Sound clinical judgment in nursing: A concept analysis. *Nursing Forum*, 54, 102–110. doi:10.1111/nuf.12303
- Oermann, M. H., Shellenbarger, T., & Gaberson, K. B. (2018). *Clinical teaching strategies in nursing* (5th ed.). New York, NY: Springer Publishing Company.
- Phillips, J. M., Stalter, A. M., Dolansky, M. A., & Lopez, G. M. (2016). Fostering future leadership in quality and safety in health care through systems thinking. *Journal of Professional Nursing*, 32, 15–24. doi:10.1016/j.profnurs.2015.06.003
- Quality and Safety Education for Nurses. (2019). Website. Retrieved from <http://qsen.org>
- Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, 10, 176–199. doi:10.1177/1745691615569000
- Sullivan, D. T. (2016). An introduction to curriculum development. In D. M. Billings & J. A. Halstead (Eds.), *Teaching in nursing: A guide for faculty* (5th ed., pp. 89–117). St. Louis, MO: Elsevier.
- Sullivan, M., Kiovsky, R. D., J. Mason, D., Hill, C. D., & Dukes, C. (2015). Interprofessional collaboration and education. *American Journal of Nursing*, 115, 47–54. doi:10.1097/01.Naj.0000461822.40440.58
- Suresh, P., Matthews, A., & Coyne, I. (2013). Stress and stressors in the clinical environment: A comparative study of fourth-year student nurses and newly qualified general nurses in Ireland. *Journal of Clinical Nursing*, 22, 770–779. doi:10.1111/j.1365-2702.2012.04145.x
- Victor, J., Ruppert, W., & Ballasy, S. Examining the relationships between clinical judgment, simulation performance, and clinical performance. *Nurse Educator*, 42, 236–239. doi: 10.1097/NNE.0000000000000359
- Zieber, M. P., & Williams, B. (2015). The experience of nursing students who make mistakes in clinical. *International Journal of Nursing Education Scholarship*, 12. doi:10.1515/ijnes-2014-0070

CLINICAL EVALUATION METHODS

After establishing a framework for evaluating students in clinical practice and exploring one's own values, attitudes, and biases that may influence evaluation, the teacher identifies a variety of methods for collecting data on student performance. Clinical evaluation methods are strategies for assessing students' performance in clinical practice. That practice may be with patients in hospitals and other healthcare facilities, in communities, in simulation and learning laboratories, and in virtual environments. Some evaluation methods are most appropriate for use by clinical educators or preceptors who are on site with students and can observe their performance; other evaluation methods assess students' knowledge, cognitive skills, and other competencies but do not involve direct observation of their performance.

There are many evaluation methods for use in nursing education. Some methods, such as reflective writing assignments, are most appropriate for formative evaluation, whereas others are useful for either formative or summative evaluation. In this chapter, varied strategies are presented for evaluating clinical performance.

■ Selecting Clinical Evaluation Methods

There are several factors to consider when selecting clinical evaluation methods to use in a course. First, the evaluation methods should provide information on student performance of the clinical competencies associated with the course. With the evaluation methods, the teacher collects data on performance to judge whether students are developing the clinical competencies or have achieved them by the end of the course. For many outcomes of a course, there are different strategies that can be used, thereby providing flexibility in choosing methods for evaluation. Most evaluation methods provide data on multiple outcomes. For example, a written assignment in which students compare two different data sets might relate to outcomes on assessment, analysis, and writing. In planning the evaluation for a clinical course, the teacher reviews the outcomes or competencies to be developed and decides which evaluation methods will be used for assessing them, recognizing that most methods provide information on more than one outcome or competency.

In clinical courses in nursing programs, students are evaluated typically on the outcomes of clinical practice, as identified in Exhibit 13.1. These relate to students' knowledge; use of evidence in practice; higher level thinking and clinical judgment; psychomotor, technical, and informatics competencies; communication, collaboration, and teamwork skills; values and professional behaviors; quality and safety competencies; leadership skills; responsibility; and self-assessment and development. Some of these competencies are easier to assess than others, but all aspects should be addressed in the evaluation process. Because of the breadth of competencies students need to develop, multiple strategies should be used for assessment in clinical courses.

Second, there are many different clinical evaluation strategies that might be used to assess performance. Varying the methods takes into account individual needs, abilities, and characteristics of learners. Some students may be more proficient in methods that depend on writing, whereas others may demonstrate their learning better in conferences and through discussions. Planning for multiple evaluation methods in clinical practice, as long as they are congruent with the outcomes to be evaluated, reflects these differences among students. It also avoids relying on one method, such as a rating scale, for determining the entire clinical grade.

Third, the teacher should always select evaluation methods that are realistic considering the number of students to be evaluated and practice opportunities (in the clinical setting or simulation). Planning for an evaluation method that depends on patients with specific health problems or particular clinical situations is not realistic considering the types of experiences with actual patients available to students. For that goal, a simulation or use of standardized patients would be more appropriate. Some methods are not feasible because of the number of students who would need to use them within the time frame of the course. Others may be too costly or require resources not available in the nursing education program or healthcare setting.

Fourth, evaluation methods can be used for either formative or summative evaluation. In the process of deciding how to evaluate students' clinical performance, the teacher should identify whether the methods will be used to provide feedback to learners (formative) or for grading (summative). With formative clinical evaluation, the focus is on the progression of students in meeting the learning goals. At the end of the practicum, course, or semester, summative evaluation establishes whether the student met those goals and is competent (Oermann, 2016). In clinical practice, students should know ahead of time whether the assessment by the teacher is for formative or summative purposes. Some of the methods designed for clinical evaluation provide feedback to students on areas for improvement and should not be graded. Other methods, such as rating scales and written assignments, can be used for summative purposes and therefore can be computed as part of the course or clinical grade.

Fifth, before finalizing the protocol for evaluating clinical performance in a course, the teacher should review the purpose of each assignment completed by students in

clinical practice and should decide on how many assignments will be required in the course. What are the purposes of these assignments, and how many are needed to demonstrate competency? In some clinical courses, students complete an excessive number of written assignments. How many assignments, regardless of whether they are for formative or summative purposes, are needed to meet the outcomes of the course? Students benefit from continuous feedback from the teacher, not from repetitive assignments that contribute little to their development of clinical knowledge and skills. Rather than daily or weekly care plans or other assignments, which may not even be consistent with current practice, once students develop the competencies, they can progress to other, more relevant learning activities.

Sixth, in deciding how to evaluate clinical performance, the teacher should consider the time needed to complete the evaluation, provide feedback, and grade the assignment. Instead of requiring a series of written assignments in a clinical course, the same outcomes might be met through discussions with students, case analysis by students in postclinical conference, group writing activities, and other methods requiring less teacher time that accomplish the same purposes. Considering the demands on nurse educators, it is important to consider one's own time when planning how to evaluate students' performance in clinical practice.

The rest of the chapter presents clinical evaluation methods for use in nursing education programs. Some of these methods, such as written assignments, were examined in earlier chapters.

■ Observation

The predominant strategy used to evaluate clinical performance is observing students in clinical practice, simulation and learning laboratories, and other settings. Although observation is widely used, there are threats to its validity and reliability. First, observations of students may be influenced by the teacher's values, attitudes, and biases, as discussed in Chapter 13, Clinical Evaluation Process. Altmiller (2016) emphasized that feedback about performance should be an unbiased reflection of observations and events. There also may be overreliance on first impressions, which might change as the teacher or preceptor observes the student over a period of time and in different situations. In any performance assessment, there needs to be a series of observations made before drawing conclusions about performance.

Second, in observing performance, there are many aspects of that performance on which the teacher may focus attention. For example, while observing a student administer an intravenous (IV) medication, the teacher may focus mainly on the technique used for its administration, ask limited questions about the purpose of the medication, and make no observations of how the student interacts with the patient. Another teacher observing this same student may focus on those other aspects. The same practice situation, therefore, may yield different observations.

Third, the teacher may arrive at incorrect judgments about the observation, such as inferring that a student is inattentive during conference when in fact the student is thinking about the comments made by others in the group. It is important to discuss observations with students, obtain their perceptions of their behavior, and be willing to modify one's own inferences when new data are presented. In discussing observations and impressions with students, the teacher can learn about their perceptions of performance; this, in turn, may provide additional information that influences the teacher's judgment about competencies.

Fourth, every observation in the clinical setting reflects only a sampling of the learner's performance during a clinical activity. An observation of the same student at another time may reveal a different level of performance. The same holds true for observations of the teacher; on some clinical days and for some classes, the teacher's behaviors do not represent a typical level of performance. An observation of the same teacher during another clinical activity and class may reveal a different quality of teaching.

Finally, similar to other clinical evaluation methods, the outcomes or competencies guide the teacher on *what* to observe. They help the teacher focus the observations of performance. All observations should be shared with the students.

Notes About Performance

It is difficult if not impossible to remember the observations made of each student for each clinical activity. For this reason, teachers need a strategy to help them remember their observations and the context in which the performance occurred. There are several ways of recording observations of students in clinical settings, simulation and learning laboratories, and other settings, such as notes about performance, checklists, and rating scales. These are summarized in Table 14.1.

The teacher can make notes that describe the observations made of students in the clinical setting; these are sometimes called *anecdotal notes*. Some teachers include only a description of the observed performance and then, after a series of observations, review the pattern of the student's performance, whereas others include a judgment or impression with each observation. Notes about observations of performance should be recorded as close to the time of the observation as possible; otherwise, it is difficult to remember what was observed and the context, for example, the patient and clinical situation, of that observation. In a study of nursing programs in both the United States and Canada, 92.4% of clinical faculty who were teaching prelicensure students used anecdotal notes for keeping records of their observations of students (Hall, 2013). Notes can be handwritten or recorded on smartphones, tablets, or other types of portable devices, and then shared with students.

Notes should be shared with students as frequently as possible; otherwise, they are not effective for feedback. In a study by Quance (2016), students reported that they preferred to have this feedback before their next clinical experience.

TABLE 14.1 Methods for Recording Observations

Notes about performance	Used for recording descriptions of observations made of students in the clinical setting, simulation, and other learning activities in which clinical nurse educators, preceptors, and others observe performance. May also include interpretations or conclusions about the performance. Often referred to as <i>anecdotal notes</i> .
Checklists	Used primarily for recording observations of specific competencies, procedures, and skills performed by students; includes list of behaviors to demonstrate competency and steps for carrying out the procedure or skill. May also include errors in performance to check.
Rating scales	Used for recording judgments about students' performance in clinical practice. Includes a set of defined clinical outcomes or competencies and scale for rating the degree of competence (with multiple levels or pass-fail).

Considering the issues associated with observations of clinical performance, the teacher should discuss observations with the students and be willing to incorporate their own judgments about the performance. Notes about performance also are useful in conferences with students, for example, at midterm and end of the term, as a way of reviewing a pattern of performance over time. When there are sufficient observations about performance, the notes can serve as documentation for ratings on the clinical evaluation tool.

Checklists

A checklist is a list of specific behaviors or actions to be observed with a place for marking whether or not they were present during the performance (Brookhart & Nitko, 2019). A checklist often lists the steps to be followed in performing a procedure or demonstrating a skill. Some checklists also include errors in performance that are commonly made. Checklists not only facilitate the teacher's observations, but they also provide a way for learners to assess their own performance. With checklists, learners can review and evaluate their performance prior to assessment by the teacher.

Checklists are used frequently in healthcare settings to assess skills of nurses and document their continuing competence in performing them. They also are used to assess performance in simulations. Many checklists and tools have been developed for evaluating the performance of students, nurses, and other health professionals in simulations. When skills are assessed in an objective structured clinical examination (OSCE) or by using standardized patients, checklists are often included to guide observations of performance of those skills.

For common procedures and skills, teachers can find checklists already prepared that can be used for evaluation, and some nursing textbooks have accompanying skills checklists. When these resources are not available, teachers can develop their own

checklists. However, there should be some consistency among faculty members in expectations of skill performance and in the checklist used for evaluation. Kardong-Edgren and Mulcock (2016) found multiple variations of skills checklists, instructional practices, and expectations of performance for their sample skill (inserting a Foley catheter) across the prelicensure program, making it difficult for students to learn this skill. Initially, it is important to review the procedure or competency to understand the steps in the procedure and critical elements in its performance. The checklist should list the steps in order and should include a scale to designate whether the student completed each step using the correct procedure. Generally a yes–no scale is used.

In designing checklists, it is important not to include every possible step, which makes the checklist too cumbersome to use, but to focus instead on critical actions and where they fit into the sequence. The goal is for students to learn how to perform a procedure and use technology safely. When there are different ways of performing a procedure, the students should be allowed that flexibility when evaluated. Exhibit 14.1 provides an example of a checklist.

EXHIBIT 14.1

SAMPLE CHECKLIST

Student Name _____

Instructions to teacher: Observe the student performing the following procedure and check the steps completed properly by the student. Check only those steps that the student performed properly. After completing the checklist, discuss performance with the student, reviewing aspects of the procedure to be improved.

IV Medication Administration

Checklist:

- ☐ Checks provider's order.
- ☐ Checks medication administration record.
- ☐ Adheres to rights of medication administration.
- ☐ Assembles appropriate equipment.
- ☐ Checks compatibility with existing IV if present.
- ☐ Explains procedure to patient.
- ☐ Positions patient appropriately.
- ☐ Checks patency of administration port or line.
- ☐ Administers medication at proper rate and concentration.
- ☐ Monitors patient response.
- ☐ Flushes tubing as necessary.
- ☐ Documents IV medication correctly.

IV, intravenous.

Rating Scales

Rating scales, also referred to as *clinical evaluation tools* or *instruments*, provide a means of recording judgments about the observed performance of students in clinical practice. A rating scale has two parts: (a) a list of outcomes or competencies the student is to demonstrate in clinical practice and (b) a scale for rating the student's performance of them.

Rating scales are most useful for summative evaluation of performance; after observing students over a period of time, the teacher arrives at conclusions about performance, rating it according to the scale provided with the tool. They also may be used to evaluate specific activities that the students complete in clinical practice, for example, rating a student's presentation of a case in clinical conference. Other uses of rating scales are to (a) help students focus their attention on important competencies to be developed, (b) give specific feedback to students about their performance, and (c) demonstrate growth in clinical competencies over a designated time period if the same rating scale is used. Rating scales also are used to assess performance in simulations. In simulations the goal of the assessment is generally formative, providing feedback to students on their judgments and actions taken in the simulation. However, simulations can also be used for high-stakes evaluation, determining students' achievement of end-of-program competencies (Bensfield, Olech, & Horsley, 2012; Oermann, Kardong-Edgren, & Rizzolo, 2016; Rizzolo, Kardong-Edgren, Oermann, & Jeffries, 2015). Using simulations for evaluation that is high stakes (students must pass the simulation to successfully pass the course) is described in Chapter 15, *Simulation and Objective Structured Clinical Examinations for Assessment*.

The same rating scale can be used for both a midterm evaluation (documenting students' progress in developing the competencies) and the final evaluation (documenting that they can safely perform them). Exhibit 14.2 shows sample competencies from a rating scale that can be used midway through a course and for the final evaluation.

Types of Rating Scales

Many types of rating scales are used for evaluating clinical performance. The scales may have multiple levels for rating performance, such as 1 to 5 or exceptional to below average, or have two levels, such as pass–fail or satisfactory–unsatisfactory. Types of scales with multiple levels for rating performance include

- Letters: A, B, C, D, E or A, B, C, D, F
- Numbers: 1, 2, 3, 4, 5
- Qualitative labels: *Excellent, very good, good, fair, and poor; exceptional, above average, average, and below average*
- Frequency labels: *Always, often, sometimes, and never*

EXHIBIT 14.2**SAMPLE COMPETENCIES FROM RATING SCALE**

Student Name _____ Faculty Name _____ Date _____

	MIDTERM			FINAL	
COMPETENCIES	S	NI	U	S	U
1. Provides patient-centered care for patients with chronic health problems across the life span					
A. Completes a comprehensive assessment using multiple sources of data					
B. Develops an individualized plan of care reflecting patient values, preferences, and needs					
C. Implements nursing interventions based on evidence					
D. Evaluates the outcomes of care					
E. Demonstrates caring behaviors					
2. Collaborates with nurses, the interprofessional team, and others in the healthcare system and community					
A. Demonstrates effective communication skills (with patients, families/caregivers, nurses, and other healthcare providers)					
B. Communicates relevant information about the patient and clinical situation using SBAR					
C. Collaborates with members of intra- and interprofessional teams					
D. Identifies resources for patients and caregivers for discharge and transitions of care					
3. . . .					

NI, needs improvement; S, satisfactory; SBAR, situation–background–assessment–recommendation; U, unsatisfactory.

Some instruments for rating clinical performance combine different types of scales, for example, rating performance of competencies on a scale of 1 to 4 based on the students' independence in practice and their knowledge, skills, and attitudes. In one school of nursing, a grade is then generated from the ratings (Altmiller, 2017).

A short description included with the letters, numbers, and labels for each of the outcomes or competencies rated improves objectivity and consistency

(Brookhart & Nitko, 2019). For example, if the tool uses a scale with numbers, short descriptions should be written to clarify the performance expected at each level. For the competency “Collects relevant data from patient,” the descriptors might be:

- 4: Differentiates relevant from irrelevant data, analyzes multiple sources of data, establishes comprehensive database, identifies data needed for evaluating all possible patient problems.
- 3: Collects significant data from patients, uses multiple sources of data as part of assessment, identifies possible patient problems based on the data.
- 2: Collects significant data from patients, uses data to develop main patient problems.
- 1: Does not collect significant data and misses important cues in data; unable to explain relevance of data for patient problems.

Many rating scales for clinical evaluation have only two levels: pass–fail or satisfactory–unsatisfactory. A survey of nursing faculty from all types of programs indicated that most faculty members ($n = 1,116$; 83%) used pass–fail or satisfactory–unsatisfactory in their clinical courses (Oermann, Saewert, Charasika, & Yarbrough, 2009). It is generally easier and more reliable for teachers to rate performance as either satisfactory or unsatisfactory (or pass–fail) rather than differentiating performance according to 4 or 5 levels of proficiency.

Any rating form used in a nursing education program or healthcare system must be clear to all stakeholders. Students, educators, preceptors, and others need to understand the meaning of the competencies and scale levels. They also need to be able to determine examples of clinical performance that reflect each level in the scale. For example, what is satisfactory and unsatisfactory performance in establishing relationships with team members? If a scale with 5 levels is used, what are the differences in establishing relationships with team members at each of those levels? All too often the meaning of the competencies in the tool and levels used to rate observed performance are not fully understood by the clinical educators using it.

Teachers should be prepared for use of the form through faculty development. The meaning of the competencies and examples of performance that reflect each level should be discussed by all educators who will be using the form. There needs to be agreement on the meaning of each of the competencies and the behaviors that represent acceptable performance of them. Without these discussions, there may be wide variability in the interpretation of the competencies and behaviors that represent a pass or fail, or a 4, 3, 2, or 1 level of performance. In addition to these discussions, teachers can practice using the form to evaluate performance of students in digitally recorded simulations.

Along with teacher preparation for using the clinical evaluation tool, some schools develop guidelines that accompany the tool to improve consistency in its use among clinical educators. Walsh, Jairath, Paterson, and Grandjean (2010) reported on the development of their Clinical Performance Evaluation Tool (CPET), based on the Quality and Safety Education for Nurses (QSEN) competencies. The CPET has three parts: (a) a one-page checklist for teachers to evaluate student performance related to the QSEN competencies, (b) a key that explains the application of the competencies to the specific clinical course, and (c) guidelines for grading performance.

Issues With Rating Scales

One problem in using rating scales with multiple levels is consistency among clinical teachers and others in determining the level of performance based on the scale. This problem can occur even when descriptions are provided for each level of the rating scale. Teachers may differ in their judgments of whether the student collected *relevant* data, whether *multiple* sources of data were used, whether the database was *comprehensive*, whether *all possible* patient problems were considered, and so forth. Scales based on frequency labels are often difficult to implement because of limited opportunities for students to practice and demonstrate a level of skill rated as *always*, *often*, *sometimes*, and *never*. How should teachers rate students' performance in situations in which they practiced the skill perhaps once or twice? Even with two-dimensional scales such as pass–fail, there is room for variability among educators.

Brookhart and Nitko (2019) identified errors that evaluators can make when using rating scales. Three of these can occur with tools that have multiple points on the scale for rating performance, such as 1 to 4:

1. *Leniency error* results when the teacher tends to rate all students toward the high end of the scale.
2. *Severity error* is the opposite of leniency, tending to rate all students toward the low end of the scale.
3. *Central tendency error* is hesitancy to mark either end of the rating scale and instead use only the midpoint of the scale. Rating students only at the extremes or only at the midpoint of the scale limits the validity of the ratings for all students and introduces the teacher's own biases into the evaluation (Brookhart & Nitko, 2019).

Three other errors that can occur with any type of clinical performance rating scale are a halo effect, personal bias, and a logical error:

4. *Halo effect* is a judgment based on a general impression of the student. With this error, the teacher lets an overall impression of the student influence the ratings of specific aspects of the student's performance. This impression is considered to create a "halo" around the student that affects the teacher's

ability to objectively evaluate and rate specific competencies on the tool. This halo may be positive, giving the student a higher rating than is deserved, or negative, letting a general negative impression of the student result in lower ratings of specific aspects of the performance.

5. *Personal bias* occurs when the teacher's biases influence ratings such as favoring nursing students who do not work while attending school over those who are employed while attending school.
6. *Logical error* results when similar ratings are given for items on the scale that are logically related to one another. This is a problem with rating scales in nursing that are too long and often too detailed. For example, there may be multiple competencies related to communication skills to be rated. The teacher observes some of these competencies but not all of them. In completing the clinical evaluation form, the teacher gives the same rating to all competencies related to communication on the tool. When this occurs, often some of the items on the rating scale can be combined.

Two other errors that can occur with performance ratings are rater drift and reliability decay (Brookhart & Nitko, 2019):

7. *Rater drift* can occur when teachers redefine the performance behaviors to be observed and assessed. Initially in developing a clinical evaluation form, teachers agree on the competencies to be rated and the scale to be used. However, over a period of time, educators may interpret them differently, drifting away from the original intent. For this reason, faculty members, clinical educators, and others involved in the clinical evaluation should discuss as a group each competency on their evaluation tool at the beginning of the course and at the midpoint. This discussion should include the meaning of the competency and what a student's performance would "look like" at each rating level in the tool. Simulated experiences in observing a performance, rating it with the tool, and discussing the rationale for the rating are valuable to prevent rater drift as the course progresses.
8. *Reliability decay* is a similar issue that can occur. Brookhart and Nitko (2019) indicated that immediately following training on using a performance rating tool, educators tend to use the tool consistently across students and with each other. As the course continues, though, faculty members may become less consistent in their ratings. Discussion of the clinical evaluation tool among course faculty, as indicated earlier, may improve consistency in use of the tool.

Although there are issues with rating scales, they remain an important clinical evaluation method because they allow teachers, preceptors, and others to rate performance over time and to note patterns of performance. Exhibit 14.3 provides guidelines for using rating scales for clinical evaluation in nursing.

EXHIBIT 14.3**GUIDELINES FOR USING RATING SCALES FOR CLINICAL EVALUATION**

1. Be alert to the possible influence of your own values, attitudes, beliefs, and biases in observing performance and drawing conclusions about it.
2. Use the clinical outcomes or competencies to focus your observations. Give students feedback on other observations made about their performance.
3. Collect sufficient data on students' performance before drawing conclusions about it.
4. Observe the student more than one time before rating performance. Rating scales, when used for clinical evaluation, should represent a *pattern* of the student's performance over a period of time.
5. If possible, observe students' performance in different clinical situations, either in the patient care setting or simulation. When not possible, develop additional strategies for evaluation so that performance is evaluated with different methods and at different times.
6. Do not rely on first impressions; they may not be accurate.
7. Always discuss observations with students, obtain their perceptions of performance, and be willing to modify your own judgments and ratings when new data are presented.
8. Review the available clinical learning activities and opportunities in the simulation and learning laboratories. Are they providing sufficient data for completing the rating scale? If not, new learning activities may need to be developed, or the competencies on the tool may need to be modified to be more realistic considering the clinical teaching circumstances.
9. Avoid using rating scales as the only source of data about a student's performance—use multiple evaluation methods for clinical practice.
10. Rate each competency individually based on the observations made of performance and conclusions drawn. If you have insufficient information about achievement of a particular competency, do not rate it—leave it blank.
11. Do not rate all students high, low, or in the middle; similarly, do not let your general impression of the student or personal biases influence the ratings.
12. If the rating form is ineffective for judging student performance, then revise and reevaluate it. Consider these questions: Does use of the form yield data that can be used to make valid decisions about students' competence? Does it yield reliable, stable data? Is it easy to use? Is it realistic for the types of learning activities students complete and that are available in clinical and simulation settings?
13. Discuss as a group (with other educators and preceptors involved in the evaluation) each competency on the rating scale. Come to agreement as to the meaning of the competencies and what a student's performance would look like at each rating level in the tool. Share examples of performance, how you would rate them, and your rationale. As a group exercise observe a video clip or other simulation of a student's performance, rate it with the tool, and come to agreement as to the rating. Exercises and discussions such as these should be held before the course begins and periodically throughout to ensure reliability across teachers and settings.

■ Written Assignments

Written assignments accompanying the clinical experience are effective methods for assessing students' problem-solving, critical thinking, and clinical reasoning; their understanding of content relevant to clinical practice; and their ability to express ideas in writing. Evaluation of written assignments was described in Chapter 9, *Assessment of Written Assignments*. There are many types of written assignments appropriate for clinical evaluation. The teacher should first specify the outcomes to be evaluated with written assignments and then decide which assignments would best assess whether those outcomes were met. The final decision is how many assignments will be required in a clinical course.

Written assignments are valuable for evaluating outcomes in face-to-face, clinical practice, and distance education courses in nursing. However, they are misused when students complete the same assignments repetitively throughout a course once the outcomes have been met. At that point, students should progress to other, more challenging learning activities. Some of the written assignments might be done as a group activity in postclinical conferences, in class, or online—teachers still can assess student progress toward meeting the outcomes but with fewer demands on their time for reviewing the assignments and providing prompt feedback on them.

Journal Writing

Journals provide an opportunity for students to describe events and experiences in their clinical practice and to reflect on them. With journals, students can “think aloud” and share their feelings with teachers. Journals are not intended to develop students' writing skills; instead, they provide a means of expressing feelings and reflections on clinical practice and other types of learning experiences, and engaging in a dialogue with the teacher about them. When journals are used for reflection, students can better understand themselves and their learning needs, build on their strengths, and identify areas for improvement (Fernandez-Pena et al., 2016). Journals also may be a strategy for assisting students to develop clinical judgment (Bussard, 2015; Lasater, 2011). Other outcomes of using journals are connecting theory and practice, assessing own strengths and weaknesses, and integrating new ideas. Issues, however, are the amount of time needed for reflection and for faculty comments, and the need to have trust between students and the teacher (Langley & Brown, 2010). Journals can be submitted in electronic formats, but should be password protected. Electronic submission of journals makes it easier for teachers to provide prompt feedback and engage in dialogue with learners, and it simplifies storing the journals.

When journals are used in a clinical course, students need to be clear about the objectives—what are the purposes of the journal? For example, a journal intended for reflection in practice would require different entries than one used for documenting

events and activities in the clinical setting as a means of communicating them to faculty. Students also need written guidelines for journal entries, including how many and what types of entries to make. Depending on the outcomes, journals may be done throughout a clinical course or at periodic intervals. Regardless of the frequency, students need immediate and meaningful feedback about their reflections and entries.

One of the issues in using journals is whether they should be graded or used solely for reflection and growth. For those educators who support grading journals, a number of strategies have been used, such as:

- Indicating a grade based on the number of journals submitted rather than on the comments and reflections in them
- Grading papers written from the journals
- Incorporating journals as part of portfolios, which then are graded
- Having students evaluate their own journals based on preset criteria developed by the students themselves
- Requiring a journal as one component among others for passing a clinical course

There are some teachers who grade the entries of a journal similar to other written assignments. However, when the purpose of the journal is to reflect on experiences in clinical practice and on the students' own behaviors, beliefs, and values, journals should not be graded. By grading journals, the teacher inhibits the student's reflection and dialogue about feelings and perceptions of clinical experiences.

Nursing Care Plans

Nursing care plans enable the student to learn the components of the nursing process and how to use the literature and other resources for writing the plan. However, a linear type of care plan does not help students learn how problems interrelate or gain higher level thinking skills. McDonald, Neumeier, and Olver (2018) suggested that linear care plans do not support students in learning how to think critically or in developing their clinical reasoning and other higher ways of thinking. If care plans are used for clinical evaluation, teachers should be cautious about the number of plans required in a course and the outcomes of such an assignment. Short assignments in which students analyze data, examine competing diagnoses and patient problems, evaluate different interventions and their evidence for practice, suggest alternative approaches, and evaluate outcomes of care are more effective than a care plan that students often paraphrase from their textbooks.

Concept Maps

Concept maps are tools used to visually display relationships among concepts. An example is provided in Figure 14.1. With a concept map, students can develop their

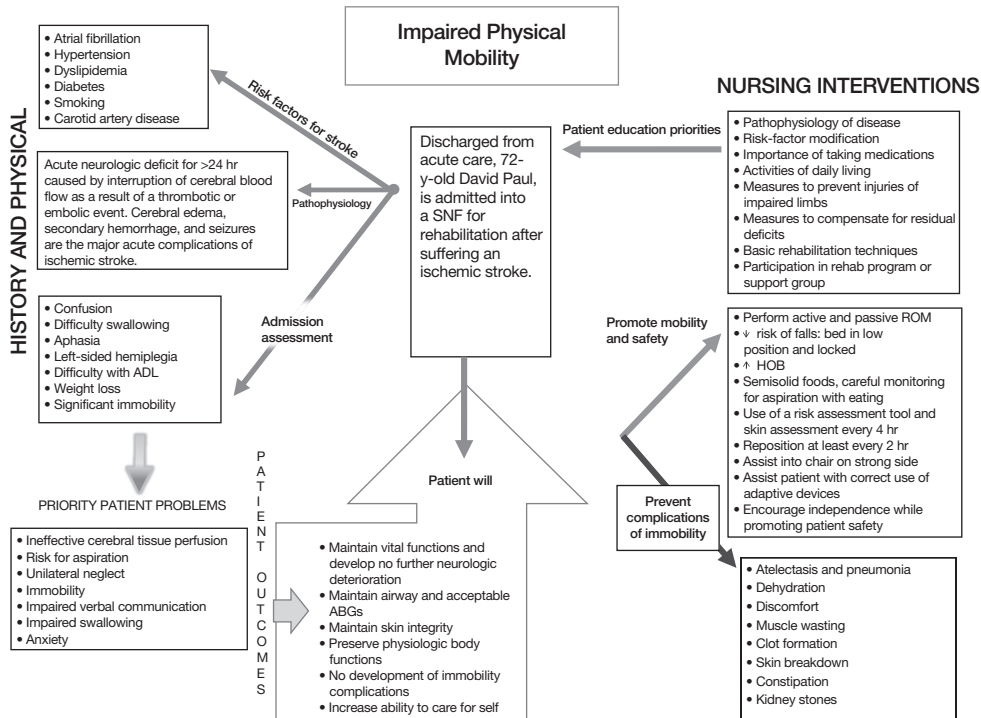


FIGURE 14.1 An example of a concept map.

ABG, arterial blood gas; ADL, activities of daily living; HOB, head of bed; ROM, range of motion; SNF, skilled nursing facility.

Source: Developed by Deanne Blach, MSN, RN. Reprinted by permission of Deanne Blach (2019).

understanding of how concepts relate to one another and visualize relationships and links (Spencer, Anderson, & Ellis, 2013). Concept maps promote students' learning, critical thinking and clinical decision-making, creativity, and ability to link theory to clinical practice (Aein & Aliakbari, 2017; Chan, 2017; Daley, Morgan, & Black, 2016; Kaddoura, Van-Dyke, & Yang, 2016). They also help students organize data as they plan for their clinical practicum; the map can be developed in a preclinical conference and then revised as the student cares for the patient. With a concept map, students can "see" graphically how assessment data, patient problems, interventions, and other aspects of care are related to one another. McDonald et al. (2018) adapted a nursing care plan and concept map into what they call a *Concepto-Plan*. With this teaching strategy, students focus on the relationships among data and nursing diagnoses.

In most cases, concept maps are best used for formative evaluation. However, with criteria established for evaluation, they also can be graded. For example, in a paper students could explain the interrelationships among concepts in the map and provide support from the literature. Other areas to assess in a concept map for patient care, depending on the goal of the assignment, are whether the assessment data are comprehensive, data are linked with the correct diagnoses and problems, nursing

interventions and treatments are relevant and based on evidence, and relationships among the concepts are indicated and accurate.

Cases, Unfolding Cases, and Case Studies

Cases, unfolding cases, and case studies were described in Chapter 7, Assessment of Higher Level Learning, because they are strategies for assessing problem-solving, clinical judgment, and higher level learning. Cases that require application of knowledge from readings and the classroom or an online component of the course can be developed for analysis by students. The scenarios can focus on patients, families, communities, the healthcare system, and other clinical situations that students might encounter in their clinical practice.

Although these assignments may be completed as individual activities, they are also appropriate for group work. Cases may be presented for group discussion and peer review in postclinical conferences and for online discussions. In online courses, the case scenario can be presented with open-ended questions and, based on student responses, other questions can be introduced for discussion. Using this approach, cases are effective for encouraging higher level thinking. By discussing cases as a clinical group, students are exposed to other possible approaches and perspectives that they may not have identified themselves. With this method, the teacher can provide feedback on the content and thought process used by students to arrive at their answers.

One advantage of short cases, unfolding cases, and case studies is that they can be graded. By using the principles described for scoring essay tests, the teacher can establish criteria for grading, develop a rubric, and score responses to the questions with the case. Otherwise, cases are useful for formative evaluation and student self-assessment.

Papers

Short papers for assessing critical thinking and other cognitive skills were described in Chapter 9, Assessment of Written Assignments. In short papers about clinical practice, students can:

- Given a data set, identify patient problems and what additional data need to be collected.
- Compare data and problems of patients for whom they have provided nursing care, identifying similarities and differences.
- Given a hypothetical patient, population, or healthcare system problem, identify possible interventions with supporting evidence.
- Select a patient, family, community, or system problem, and describe relevant interventions with evidence for their use.
- Identify one intervention they used and identify alternative approaches with supporting evidence.

- Identify a decision made in clinical practice involving patients or staff, describe why they made that decision, and propose one other approach that could be used with rationale.
- Identify a problem or an issue they had in clinical practice, critique the approaches they used for resolving it, and identify alternate approaches.

Short written assignments in clinical courses may be more beneficial than longer assignments because with long papers, students often summarize from the textbook and other literature without engaging in any of their own thinking about the content. Short papers can be used for formative evaluation or graded.

Term papers also may be written about clinical practice. With term papers, students can critique and synthesize relevant literature and write a paper about how that literature relates to patient care. Or they might prepare a paper on the use of selected concepts in patient care. If the term paper includes the submission of drafts combined with prompt feedback on writing from the teacher, it can be used as a strategy for improving writing skills. Although drafts of papers are assessed but not graded, the final product is graded by the teacher.

There are many other written assignments that can be used for clinical evaluation in a nursing course. Similar to any assignment in a course, requirements for papers should be carefully thought out: What outcomes will be met with the assignment, how will they contribute to clinical evaluation in the course, and how many of those assignments does a student need to complete for competency? In planning the clinical evaluation protocol, the teacher should exercise caution in the type and number of written assignments so that they promote learning without unnecessary repetition. Guidelines for evaluating written assignments were presented in Chapter 9, Assessment of Written Assignments, and therefore are not repeated here.

■ Electronic Portfolio

An *electronic portfolio* (e-portfolio) is a collection of projects and materials (also referred to as artifacts) developed by the student that documents achievement of the outcomes of the clinical course. With a portfolio, students can demonstrate what they have learned in clinical practice and the competencies they have developed. Portfolios are valuable for clinical evaluation because students provide evidence in their portfolios to confirm their clinical competencies and document new learning and skills acquired in a course. The e-portfolio can include evidence of student learning for a series of clinical experiences or over the duration of a clinical course. Most portfolios are developed electronically, which facilitates updating and revising entries, as compared with portfolios that include hard copies of materials. In addition to easy updating, prior versions of the e-portfolio can be archived.

E-portfolios are used increasingly in doctor of nursing practice programs to provide documentation of student achievement of competencies and program outcomes (Anderson, DesLauriers, Horvath, Slota, & Farley, 2017; Melander, Hampton, Hardin-Pierce, & Ossege, 2018; Moriber et al., 2014; Willmarth-Stec & Beery, 2015). Other nursing education programs are using e-portfolios in prelicensure, RN-to-BSN, and master's programs. E-portfolios provide a way of assessing cognitive, reflective, and affective skills of students and for nurses to document their achievements, experiences, and own development (Green, Wyllie, & Jackson, 2014). Portfolios can be evaluated and graded by faculty members based on predetermined criteria. They also can be used for students' self-assessment of their progress in meeting personal and professional goals and not be evaluated by the teacher.

Brookhart and Nitko (2019) identified two types of portfolios: best work and growth. Best-work portfolios contain the students' best final products. These provide evidence that the students have demonstrated certain competencies and achievements in clinical practice, and thus are appropriate for summative clinical evaluation. Growth portfolios are designed for monitoring students' progress and for self-reflection on learning outcomes at several points in time. These contain products and work of the students in process and at the intermediate stages for the teacher to review and provide feedback (Brookhart & Nitko, 2019).

For clinical evaluation, these purposes can be combined. The portfolio can be developed initially for growth and learning, with products reviewed periodically by the teacher for formative evaluation. The portfolio can then be submitted as a best-work portfolio with completed products providing evidence of clinical competencies. The best-work portfolio can be graded.

The contents of the portfolio depend on the outcomes of the clinical course and competencies to be developed. Many types of materials and documentation can be included in a portfolio. For example, students can include papers they completed in the course, reports of group work, reports and analyses of observations made in the clinical setting, self-reflections on clinical experiences, concept maps, and other products they developed in their clinical practice. The key is for students to choose documents that demonstrate their learning and development of clinical competencies. By assessing the portfolio, the teacher should be able to determine whether the students met the outcomes of the course.

There are several steps to follow in using e-portfolios for clinical evaluation. Brookhart and Nitko (2019) suggested steps for crafting a portfolio system, some of which are relevant for nursing courses.

Step 1: Identify the purpose of the portfolio.

- Why is an e-portfolio useful in the course? What goals will it serve?
- Will the e-portfolio serve as a means of assessing students' development of clinical competencies, focusing predominantly on the growth of the students?

Will the portfolio provide evidence of the students' best work in clinical practice, including documents that reflect their learning over a period of time? Or, will the e-portfolio meet both demands, enabling the teacher to give feedback to students on the process of learning and projects on which they are working, as well as providing evidence of their accomplishments and achievements in clinical practice?

- Will the e-portfolio be used for formative or summative evaluation? Or both?
- Will the e-portfolio provide assessment data for use in a clinical course? Or will it be used for program evaluation and accreditation?
- Will the portfolio serve as a means of assessing prior learning and therefore have an impact on the types of learning activities or courses that students complete, for example, for assessing the prior learning of RNs entering a higher degree program or for licensed practical nurses entering an associate degree program?
- What is the role of the students, if any, in defining the focus and content of the e-portfolio?

Step 2: Identify the type of documents and content to be included in the e-portfolio.

- What types of documents or artifacts are required in the e-portfolio, for example, products developed by students, descriptions of projects in which the students were involved, descriptions of clinical learning activities and reflections, observations made in clinical practice and analysis of them, and papers completed by the students, among others?
- In addition to required entries, what other types of content might be included in the e-portfolio?
- Who determines the content of the e-portfolio and the types of documents? Teacher only? Student only? Or both?
- Will the type of documents be the same for all students or individualized by the student?
- What is the minimum number of documents to be considered satisfactory?
- How should the e-portfolio be organized, or will the students choose how to organize the materials included in the portfolio?
- When should the e-portfolio be submitted to the teacher for review and feedback? Or for final evaluation and grading?
- Will the teacher and student meet in a conference to discuss the e-portfolio?

Step 3: Decide on the evaluation of the e-portfolio, including criteria for evaluation of individual artifacts and the portfolio overall.

- How will the e-portfolio be integrated within the clinical evaluation grade and course grade, if at all? Or is the e-portfolio to be used for program evaluation and accreditation instead of a course?
- What criteria will be used to evaluate, and perhaps score, each document included in the e-portfolio and the portfolio as a whole?
- Will only the teacher evaluate the e-portfolio and its entries? Will only the students evaluate their own progress and work? Or will the evaluation be a collaborative effort?
- Are there rubrics available for scoring the e-portfolio and individual document? If not, who will develop the rubrics?

These steps and questions to be answered provide guidelines for nurse educators in incorporating an e-portfolio for clinical evaluation in a course or for other purposes in the nursing education program.

■ Conferences

The ability to present ideas orally is an important outcome of clinical practice. Sharing information about a patient, leading others in discussions about clinical practice, presenting ideas in a group format, and giving various types of presentations are skills that students need to develop in a nursing program. Working with nursing staff members, other providers, the healthcare team, and community members requires the ability to communicate effectively with varied individuals and groups. Conferences provide a method for developing oral communication skills and for evaluating competency in this area. Discussions also lead to problem-solving and higher level thinking if questions are open ended and geared to these outcomes, as discussed in Chapter 7, Assessment of Higher Level Learning.

Many types of conferences are appropriate for clinical evaluation, depending on the outcomes to be met. Preclinical conferences take place prior to beginning a clinical learning activity and allow students to clarify their understanding of patient problems, interventions, and other aspects of clinical practice. Similar types of conferences can be held prior to a simulation. In these conferences, the teacher can assess students' knowledge and provide feedback to them. Postclinical conferences, held at the end of a clinical learning activity or at a predetermined time during the clinical practicum, provide an opportunity for the teacher to assess students' abilities to use concepts and knowledge in patient care, plan care and think through possible approaches to use, assess the outcomes of interventions, problem solve and think critically, collaborate with others, and achieve other outcomes, depending on the focus of the discussion. In clinical conferences, students also can examine ethical

dilemmas; cultural aspects of care; and issues facing patients, families, communities, providers, and the healthcare system. In discussions such as these, students can examine different perspectives and approaches that could be taken.

Although many clinical conferences will be face-to-face with the clinical nurse educator or preceptor on site with the students, conferences also can be conducted online (Berkstresser, 2016; Hannans, 2019). Online conferences can be asynchronous, held a few days after the clinical practicum to give students time to reflect on their experiences. In a study on asynchronous online postclinical conferences, Hannans (2019) reported that students benefited from time for reflection, and they were able to participate equally in the discussion.

Criteria for evaluating conferences include the ability of students to:

- Present ideas clearly and in a logical sequence to the group.
- Participate actively in the group discussion.
- Offer ideas relevant to the topic.
- Demonstrate knowledge of the content discussed in the conference.
- Offer different perspectives to the topic or share their reflections to encourage new learning among the group members.
- Assume a leadership role, if relevant, in promoting group discussion and arriving at group decisions.

Most conferences are evaluated for formative purposes, with the teacher giving feedback to students as a group or to the individual who led the group discussion. When conferences are evaluated as a portion of the clinical or course grade, the teacher should have specific criteria to guide the evaluation and should use a scoring rubric. Exhibit 14.4 provides a sample form that can be used to evaluate how well a student leads a clinical conference or to assess student participation in a conference.

Another type of discussion occurs with simulation and is referred to as *debriefing*. Debriefing is the discussion following a simulation (post-event) or during it (with-in-event), allowing students to reflect on their experiences and for the facilitator to provide feedback. Debriefing must be carried out in a safe environment. Through this discussion, students receive information about their performance (feedback) with the goals of facilitating deep learning, identifying and filling in gaps in learning, promoting student self-reflection, and improving their future performance (Dreifuerst, 2012; Fey & Jenkins, 2015; Palaganas, Fey, & Simon, 2016; Sawyer, Eppich, Brett-Fleegler, Grant, & Cheng, 2016). Debriefing can be guided by the facilitator (facilitator-guided), or the students can guide their own discussions (self-guided; Sawyer et al., 2016). Debriefing a simulation is intended for feedback and learning, not for grading.

EXHIBIT 14.4

EVALUATION OF PARTICIPATION IN CLINICAL CONFERENCE

Student's Name _____

Conference Topic _____

Date _____

Rate the behaviors by circling the appropriate number. Some behaviors will not be applicable depending on student role in the conference; mark those as NA.

BEHAVIORS	RATING					
	POOR			EXCELLENT		
States goals of conference.	1	2	3	4	5	NA
Leads group in discussion.	1	2	3	4	5	NA
Asks thought-provoking questions.	1	2	3	4	5	NA
Uses strategies that encourage all students to participate.	1	2	3	4	5	NA
Participates actively in discussion.	1	2	3	4	5	NA
Includes important content.	1	2	3	4	5	NA
Bases interventions on evidence for practice.	1	2	3	4	5	NA
Offers new perspectives to group.	1	2	3	4	5	NA
Considers different points of view.	1	2	3	4	5	NA
Assists group members in recognizing biases and values that may influence decision-making.	1	2	3	4	5	NA
Is enthusiastic about conference topic.	1	2	3	4	5	NA
Is well prepared for conference discussion.	1	2	3	4	5	NA
If leading group, monitors time.	1	2	3	4	5	NA
Develops quality materials to support discussion.	1	2	3	4	5	NA
Summarizes major points discussed at end of conference.	1	2	3	4	5	NA

NA, not applicable.

Media Clips

Media clips, short segments of a video or audio clip, may be used as a basis for discussions in postclinical conferences or critiqued by students as a clinical assignment. Media clips often are more effective than written descriptions of a scenario because they allow the student to visualize a clinical situation or listen to interactions, hear

sounds, and other types of recordings. The segment viewed or listened to by students should be short so they can focus on critical aspects related to the outcomes to be evaluated. Media clips are appropriate for assessing whether students can apply concepts and knowledge to the patient or clinical situation depicted in the media clip, observe and collect data, identify possible problems, identify priority actions and interventions, and explore one's own feelings and responses.

Students can answer questions about the media clips as part of a graded learning activity. Otherwise, media clips are valuable for formative evaluation, particularly in a group format in which students discuss their ideas and receive feedback from the teacher and their peers.

■ Group Projects

Most of the clinical evaluation methods presented in this chapter focus on individual student performance, but group projects also can be assessed as part of the clinical evaluation in a course. Some group work is short term—lasting only for the time it takes to develop a product such as a poster or group presentation. Other groups may be formed for the purpose of cooperative learning with students working in small groups or teams in clinical practice over a longer period of time. Many group projects in a clinical course are done with nursing students in that course. However, with the need to prepare healthcare professions to work in teams, some of the group projects can be interprofessional. Interprofessional group projects provide a way for students from multiple disciplines to work together as a team to meet a common goal. For example, nursing and medical students and resident and attending physicians could work on a quality-improvement project to meet an identified need on their unit.

There are different approaches for grading group projects. The same grade can be given to every student in the group, that is, a group grade, although this does not take into consideration individual student effort and contribution to the group product. Another approach is for the students to indicate in the finished product the parts they contributed, providing a way of assigning individual student grades, with or without a group grade. Students also can provide a self-assessment of how much they contributed to the group project, which can then be integrated into their grade. Alternatively, students can prepare both a group and an individual product. Rubrics should be used for assessing group projects and should be geared specifically to the project outcomes and design.

To assess students' participation and collaboration in the group, the rubric also needs to reflect the goals of group work. With small groups, the teacher can observe and rate individual student cooperation and contributions to the group. However, this is often difficult because the teacher is not a member of the group, and the group

dynamics change when the teacher is present. As another approach, students can assess the participation and cooperation of their peers. These peer evaluations can be used for the students' own development and shared among peers but not with the teacher, or can be incorporated in the grade for the group project. In one study, nursing students supported the opportunity to grade individual student contributions to a group project because it reduced the chance of some students not contributing (Shiu, Chan, Lam, Lee, & Kwong, 2012). Cook et al. (2017) developed a mathematical formula for dividing the credit allotted to group work among members of the group based on peer assessment of the contribution of each member. Students who contributed more to the group received a greater share of the credit than others. An easy-to-use form for peer evaluation of group participation is found in Exhibit 14.5. Students also can assess their own participation in the group.

■ Self-Assessment

Self-assessment is the ability of students to evaluate their own clinical competencies and identify where further learning is needed. Self-assessment begins with the first clinical course and develops throughout the nursing education program, continuing into professional practice. Through self-assessment, students examine their clinical performance and identify both strengths and areas for improvement. Using students' self-assessments, teachers can develop plans to assist students in gaining the knowledge and skills they need to meet the outcomes of the course. It is important for teachers to establish a positive climate for learning in the course, or students will not be likely to share their self-assessments with them.

In addition to developing a supportive learning environment, the teacher should hold planned conferences with each student to review performance. In these conferences, the teacher can:

- Give specific and instructional feedback on performance.
- Obtain the student's own perceptions of competencies and progress.
- Identify strengths and areas for learning from the teacher's and student's perspectives.
- Plan with the student learning activities for improving performance, which is critical if the student is not passing the clinical course.
- Enhance communication between teacher and student.

Conferences with students about performance are critical when students are not progressing in the course: these conferences and other requirements are discussed in Chapter 17, Grading.

EXHIBIT 14.5

RUBRIC FOR PEER EVALUATION OF PARTICIPATION IN GROUP PROJECT

PARTICIPATION RUBRIC					
Directions: Complete for each group member. Name _____					
SCORE	AREA ASSESSED	EXCELLENT = 4	GOOD = 3	POOR = 2	UNACCEPTABLE = 1
_____	Amount of group work	Did a full share of the work and often more than required; willingly helped others	Did an equal share of the work; did additional work when asked	Did less than an equal share of the work; rarely helped others	Did not complete all of assigned group work; never helped others
_____	Quality of group work	Contributions to the group and products met outcomes and were of high quality	Contributions to the group and products met most outcomes and were satisfactory	Some of the contributions and products did not meet outcomes and were of poor quality	Contributions to the group were limited and were consistently of poor quality
_____	Organization	Led group in getting organized; arranged meeting times and places	Helped others with organizing group; was flexible with meeting times and places	Did not contribute to organizing group; was not flexible with meeting times and places	Never participated in organizing group or arranging meeting times and places
_____	Attendance at group meetings	Attended all meetings; was on time	Attended most meetings; was on time	Missed some meetings or was consistently late	Missed many meetings; did not provide group with acceptable reasons
_____	Group participation	Provided many valuable ideas with rationale; considered views of others and was willing to modify own perspective	Provided good suggestions; considered views of others but did not readily change own perspective	Listened to group and occasionally shared own ideas	Participated minimally in discussions; ideas were often not relevant to topic
_____	Deadlines	Completed group work prior to deadline or on time	Completed most group work on time	Was late with group work; required constant reminders from group	Did not complete all of assigned group work or was always late

(continued)

EXHIBIT 14.5**RUBRIC FOR PEER EVALUATION OF PARTICIPATION IN GROUP PROJECT** (*continued*)

SCORE	AREA ASSESSED	EXCELLENT = 4	GOOD = 3	POOR = 2	UNACCEPTABLE = 1
_____	Providing and receiving feedback	Always provided specific, clear feedback to the group in a respectful manner; accepted feedback and used for revision	Provided general feedback to the group in a respectful manner; listened to feedback from others	Provided some feedback to the group; occasionally was offensive in how feedback was given; did not accept feedback from group	Did not provide any specific feedback to group; was occasionally rude when providing feedback; never listened to feedback from group
Total Score: _____					

Some students have difficulty assessing their own performance. This is a developmental process, and in the beginning of a nursing education program, students need more guidance in assessing their performance than at the end. Self-evaluation is appropriate only for formative evaluation and should not be graded.

■ Summary

This chapter built on concepts of clinical evaluation examined in Chapter 13, Clinical Evaluation Process. Many clinical evaluation methods are available for assessing student competencies in clinical practice. The teacher should choose evaluation methods that provide information on how well students are performing in the clinical setting. The teacher also decides whether the evaluation method is intended for formative or for summative evaluation. Some of the methods designed for clinical evaluation are strictly used to provide feedback to students on areas for improvement and are not graded. Other methods, such as rating forms and certain written assignments, may be used for summative purposes.

The predominant method for clinical evaluation is observing the performance of students in clinical practice. Although observation is widely used, there are threats to its validity and reliability. Observations of students may be influenced by the values, attitudes, and biases of the teacher or preceptor, as discussed in Chapter 13, Clinical Evaluation Process. In observing clinical performance, there are many aspects of that performance on which the teacher may focus attention. Every observation reflects

only a sampling of the learner's performance during a clinical learning activity. Issues such as these point to the need for a series of observations before arriving at conclusions about performance. There are several ways of recording observations of students—notes about performance, checklists, and rating scales.

There are many types of written assignments useful for clinical evaluation depending on the outcomes to be assessed: reflective journal, nursing care plan, concept map, case analysis, and a paper on some aspect of clinical practice. Written assignments can be developed as a learning activity and reviewed by the teacher and/or peers for formative evaluation, or they can be graded.

An e-portfolio is a collection of materials that students develop in clinical practice over a period of time. With a portfolio, students provide evidence to confirm their clinical competencies and document the learning that occurred in clinical practice. Other clinical evaluation methods include conferences, group projects, and self-assessment. The evaluation methods presented in this chapter provide the teacher with a wealth of methods from which to choose in evaluating students' clinical performance.

■ References

- Aein, F., & Aliakbari, F. (2017). Effectiveness of concept mapping and traditional linear nursing care plans on critical thinking skills in clinical pediatric nursing course. *Journal of Education and Health Promotion*, 6, 13. doi:10.4103/jehp.jehp_49_14
- Altmiller, G. (2016). Strategies for providing constructive feedback to students. *Nurse Educator*, 41, 118–119.
- Altmiller, G. (2017). Content validation of a Quality and Safety Education for Nurses-based clinical evaluation instrument. *Nurse Educator*, 42, 23–27. doi:10.1097/nne.0000000000000307
- Anderson, K. M., DesLauriers, P., Horvath, C. H., Slota, M., & Farley, J. N. (2017). From metacognition to practice cognition: The DNP e-portfolio to promote integrated learning. *Journal of Nursing Education*, 56, 497–500. doi:10.3928/01484834-20170712-09
- Bensfield, L. A., Olech, M. J., & Horsley, T. L. (2012). Simulation for high-stakes evaluation in nursing. *Nurse Educator*, 37, 71–74. doi:10.1097/NNE.0b013e3182461b8c
- Berkstresser, K. (2016). The use of online discussions for post-clinical conference. *Nurse Education in Practice*, 16, 27–32. doi:10.1016/j.nepr.2015.06.007
- Brookhart, S. M., & Nitko, A. J. (2019). *Educational assessment of students* (8th ed.). New York, NY: Pearson Education.
- Bussard, M. E. (2015). Clinical judgment in reflective journals of prelicensure nursing students. *Journal of Nursing Education*, 54, 36–40. doi:10.3928/01484834-20141224-05
- Chan, Z. C. Y. (2017). A qualitative study on using concept maps in problem-based learning. *Nurse Education in Practice*, 24, 70–76. doi:10.1016/j.nepr.2017.04.008
- Cook, A. R., Hartman, M., Luo, N., Sng, J., Fong, N. P., Lim, W. Y., . . . Koh, G. C. (2017). Using peer review to distribute group work marks equitably between medical students. *BMC Medical Education*, 17, 172. doi:10.1186/s12909-017-0987-z
- Daley, B. J., Morgan, S., & Black, S. B. (2016). Concept maps in nursing education: A historical literature review and research directions. *Journal of Nursing Education*, 55, 631–639. doi:10.3928/01484834-20161011-05

- Dreifuerst, K. T. (2012). Using debriefing for meaningful learning to foster development of clinical reasoning in simulation. *Journal of Nursing Education*, 51, 326–333. doi:10.3928/01484834-20120409-02
- Fernandez-Pena, R., Fuentes-Pumarola, C., Malagon-Aguilera, M. C., Bonmati-Tomas, A., Bosch-Farre, C., & Ballester-Ferrando, D. (2016). The evaluation of reflective learning from the nursing student's point of view: A mixed method approach. *Nurse Education Today*, 44, 59–65. doi:10.1016/j.nedt.2016.05.005
- Fey, M. K., & Jenkins, L. S. (2015). Debriefing practices in nursing education programs: Results from a national study. *Nursing Education Perspectives*, 36, 361–366.
- Green, J., Wyllie, A., & Jackson, D. (2014). Electronic portfolios in nursing education: A review of the literature. *Nurse Education in Practice*, 14, 4–8. doi:10.1016/j.nepr.2013.08.011
- Hall, M. A. (2013). An expanded look at evaluating clinical performance: Faculty use of anecdotal notes in the U.S. and Canada. *Nurse Education in Practice*, 13, 271–276. doi:10.1016/j.nepr.2013.02.001
- Hannans, J. (2019). Online clinical post conference: Strategies for meaningful discussion using VoiceThread. *Nurse Educator*, 44, 29–33. doi:10.1097/nne.0000000000000529
- Kaddoura, M., Van-Dyke, O., & Yang, Q. (2016). Impact of a concept map teaching approach on nursing students' critical thinking skills. *Nursing & Health Sciences*, 18, 350–354. doi:10.1111/nhs.12277
- Kardong-Edgren, S., & Mulcock, P. M. (2016). Angoff method of setting cut scores for high-stakes testing: Foley catheter checkoff as an exemplar. *Nurse Educator*, 41, 80–82. doi:10.1097/nne.0000000000000218
- Langley, M. E., & Brown, S. T. (2010). Perceptions of the use of reflective learning journals in online graduate nursing education. *Nursing Education Perspectives*, 31, 12–17.
- Lasater, K. (2011). Clinical judgment: The last frontier for evaluation. *Nurse Education in Practice*, 11, 86–92. doi:10.1016/j.nepr.2010.11.013
- McDonald, M. S., Neumeier, M., & Olver, M. E. (2018). From linear care plan through concept map to Concepto-Plan: The creation of an innovative and holistic care plan. *Nurse Education in Practice*, 31, 171–176. doi:10.1016/j.nepr.2018.05.005
- Melander, S., Hampton, D., Hardin-Pierce, M., & Ossege, J. (2018). Development of a rubric for evaluation of the DNP portfolio. *Nursing Education Perspectives*, 39, 312–314. doi:10.1097/01.Nep.0000000000000381
- Moriber, N. A., Wallace-Kazer, M., Shea, J., Grossman, S., Wheeler, K., & Conelius, J. (2014). Transforming doctoral education through the clinical electronic portfolio. *Nurse Educator*, 39, 221–226. doi:10.1097/nne.0000000000000053
- Oermann, M. H. (2016). Program evaluation: An introduction. In M. H. Oermann (Ed.), *A systematic approach to assessment and evaluation of nursing programs*. Philadelphia, PA: National League for Nursing/Wolters Kluwer.
- Oermann, M. H., Kardong-Edgren, S., & Rizzolo, M. A. (2016). Summative simulated-based assessment in nursing programs. *Journal of Nursing Education*, 55, 323–328. doi:10.3928/01484834-20160516-04
- Oermann, M. H., Saewert, K. J., Charasika, M., & Yarbrough, S. S. (2009). Assessment and grading practices in schools of nursing: National survey findings Part I. *Nursing Education Perspectives*, 30, 274–278.
- Palaganas, J. C., Fey, M., & Simon, R. (2016). Structured debriefing in simulation-based education. *AACN Advanced Critical Care*, 27, 78–85. doi:10.4037/aacnacc2016328

- Quance, M. A. (2016). Nursing students' perceptions of anecdotal notes as formative feedback. *International Journal of Nursing Education Scholarship*, 13, 20180021. doi:10.1515/ijnes-2015-0053
- Rizzolo, M. A., Kardong-Edgren, S., Oermann, M. H., & Jeffries, P. R. (2015). The National League for Nursing project to explore the use of simulation for high stakes assessment: Process, outcomes and recommendations. *Nursing Education Perspectives*, 36, 299–303.
- Sawyer, T., Eppich, W., Brett-Fleegler, M., Grant, V., & Cheng, A. (2016). More than one way to debrief: A critical review of healthcare simulation debriefing methods. *Simulation in Healthcare*, 11, 209–217. doi:10.1097/sih.0000000000000148
- Shiu, A. T. Y., Chan, C. W. H., Lam, P., Lee, J., & Kwong, A. N. L. (2012). Baccalaureate nursing students' perceptions of peer assessment of individual contributions to a group project: A case study. *Nurse Education Today*, 32, 214–218. doi:10.1016/j.nedt.2011.03.008
- Spencer, J. R., Anderson, K. M., & Ellis, K. K. (2013). Radiant thinking and the use of the mind map in nurse practitioner education. *Journal of Nursing Education*, 52, 291–293. doi:10.3928/01484834-20130328-03
- Walsh, T., Jairath, N., Paterson, M., & Grandjean, C. (2010). Quality and Safety Education for Nurses clinical evaluation tool. *Journal of Nursing Education*, 49, 517–522. doi:10.3928/01484834-20100630-06
- Willmarth-Stec, M., & Beery, T. (2015). Operationalizing the student electronic portfolio for doctoral nursing education. *Nurse Educator*, 40, 263–265. doi:10.1097/NNE.0000000000000161

SIMULATION AND OBJECTIVE STRUCTURED CLINICAL EXAMINATION FOR ASSESSMENT

Simulation is used widely for instruction in nursing. In a simulation, students can gain knowledge about patient care, develop competencies in communication and teamwork, use clinical judgment and reflect on actions taken in the scenario, and develop psychomotor and clinical skills. Simulation also is a strategy for assessment, including high-stakes evaluation, if nurse educators adhere to guidelines to ensure validity and reliability. Some simulations incorporate standardized patients, who are actors who portray the role of a patient with a specific diagnosis or condition. With standardized patients, students can be evaluated on their history and physical examination skills, communication strategies, and other competencies. Another method for evaluating skills and clinical competencies of nursing students is an objective structured clinical examination (OSCE). In an OSCE, students rotate through stations where they complete an activity or perform a skill, which then can be evaluated. This chapter examines these methods for assessing clinical competencies of students.

■ Simulations

Simulation allows learners to experience a clinical situation without the risks. With simulations, students can develop their psychomotor and technological skills and practice those skills to maintain their competence. Simulations, particularly those involving high-fidelity simulators, enable students to think through clinical situations and make independent decisions. With high-fidelity simulation and complex scenarios, students can assess a patient and clinical situation, analyze data, make decisions about priority problems and actions to take, implement those interventions, and evaluate outcomes. High-fidelity simulation can be used to guide students' development of clinical judgment skills, especially when combined with high-quality debriefing following the simulation (Chmil, Turk, Adamson, & Larew, 2015; Fey & Jenkins, 2015; Klenke-Borgmann, 2019; Lasater, 2007, 2011).

Another outcome of instruction with simulations is the opportunity to have deliberate practice of skills. Simulations allow students to practice skills, both cognitive and motor, until competent and receive immediate feedback on performance (Kardong-Edgren, Oermann, & Rizzolo, 2019; Oermann, Molloy, & Vaughn, 2014; Oermann, Muckler, & Morgan, 2016; Owen, Garbett, Coburn, & Amar, 2017; Reed, Pirotte, et al., 2016; Sullivan, 2015). Through simulations, students can develop their communication and teamwork skills and apply quality and safety guidelines to practice. Increasingly simulations are used to provide experiences for students in working with other healthcare profession students and providers (Feather, Carr, Reising, & Garletts, 2016; Furseth, Taylor, & Kim, 2016; Horsley et al., 2016; Lee, Jang, & Park, 2016; Reed, Horsley, et al., 2016; Rutherford-Hemming & Lioce, 2018).

Given the limited time for clinical practice in many programs and the complexity of skills to be developed by students, simulations are important as a clinical teaching strategy. Simulations can ease the shortage of clinical experiences for students because of clinical agency restrictions and fewer available practice hours in a curriculum. A study by the National Council of State Boards of Nursing suggested that simulation may be used as a replacement for clinical experiences (Hayden, Smiley, Alexander, Kardong-Edgren, & Jeffries, 2014).

Guidelines for Simulation-Based Assessment

Simulations not only are effective for instruction in nursing, but they also are useful for assessment. The availability of high-fidelity simulators has expanded opportunities for performance evaluation (Mitchell et al., 2018). A simulation can be developed for students to demonstrate procedures and technologies, conduct assessments, analyze data presented in a scenario, decide on priority actions to take in a situation, and evaluate the effects of their decisions. Each of these outcomes can be assessed for feedback to students or for verifying students' competencies for high-stakes evaluations. In a high-stakes evaluation, the student needs to demonstrate competency in order to pass the course or graduate from the nursing program, or for some other decision with significant consequences.

In formative assessment using simulation, the teacher, referred to as the *facilitator*, shares observations about the student's performance and other behaviors with the student. The goal of formative assessment is to provide feedback to students individually and to the team, if relevant, to guide further development of competencies. This feedback is an essential component of the facilitator's role in the simulation. In contrast, the goal of summative assessment is to determine students' competence. Summative assessment verifies that students can perform the required clinical competencies.

There are different types of simulations that can be used for assessment. Case scenarios that students analyze can be presented in paper-and-pencil format or

through multimedia. Many computer simulations are available for use in assessment. Simulations can be developed with models and manikins for evaluating skills and procedures, and for evaluation with standardized patients. With high-fidelity simulation, teachers can identify outcomes and clinical competencies to be assessed, present various clinical events and scenarios for students to analyze and then take action, and evaluate students' decisions and performance in these scenarios. Prebriefing, the introductory phase of a simulation, prepares students for learning in the simulation. Page-Cuttrara and Turk (2017) found that a structured prebriefing (with concept-mapping activities and guided reflection) improved students' competency performance, clinical judgment, and prebriefing experience. Following the simulation in the debriefing session, the students as a group can analyze the scenario and critique their actions and decisions, with facilitators (and standardized patients) providing feedback. The debriefing also promotes students' development of clinical-judgment skills (Dube et al., 2019; Fey & Jenkins, 2015; Klenke-Borgmann, 2019; Lasater, 2011; Victor, 2017). Many nursing education programs have simulation laboratories with high-fidelity and other types of simulators, clinically equipped examination rooms, manikins and models for skill practice and assessment, areas for standardized patients, and a wide range of multimedia that facilitate performance evaluations. The rooms can be equipped with two-way mirrors, video cameras, microphones, and other media for observing and performance rating by faculty and others.

In simulation-based assessment, the first task is to identify the objectives of the assessment and the knowledge and competencies to be evaluated. This is important because these guide developing the simulation and writing the scenario. If the competencies are skill oriented, use of a high-fidelity simulator may not be necessary and a model or partial task trainer, which allows students to perform a specific task such as venipuncture, can be used for assessment. In contrast, if the assessment is to determine students' ability to analyze a complex clinical situation, arrive at clinical judgments about the best approaches to take, demonstrate a range of clinical competencies, and communicate effectively, assessment with high-fidelity patient simulators would be appropriate.

Once the objectives of the assessment and the knowledge and skills to be evaluated are identified, the teacher can plan the specifics of the assessment. The assessment processes need to be defensible if high-stakes decisions will be made based on the assessment (Tavares et al., 2018). The simulation needs to focus on the intended purpose and require that students *use* the knowledge and competencies. This is a key principle for the simulation to be valid. Validity is the extent to which the simulation measures what it was intended to measure (Boulet & Murray, 2010; Oermann, Kardong-Edgren, & Rizzolo, 2016a; O'Leary, 2015). The teacher should have colleagues and other experts review the simulation to ensure that it is appropriate for

the objectives, that students would need to apply their knowledge and skills in it, and that it represents a realistic clinical scenario. This review by others helps establish the validity of the simulation. For high-stakes decisions, the simulation can be piloted with different levels of students in the nursing program. The performance of senior-level or graduate nursing students in a simulation should be different from that of beginning students. Piloting the simulation with different levels of students also may reveal issues to be resolved before using it for an assessment.

Some simulations for assessment are short, for example, a few minutes, such as when evaluating performance of skills. However, when evaluating students' competencies such as communication ability and teamwork, the simulation may last 30 minutes (Mudumbai, Gaba, Boulet, Howard, & Davies, 2012). When the aim is to provide feedback on or verify students' clinical judgment, their ability to manage a complex scenario, and other higher level skills, it is likely that longer simulations will be needed.

A key point in evaluating student performance in simulation for high-stakes and other summative decisions is the need for a tool that produces valid and reliable results. An example of a rating scale used for evaluating students in a simulation, with demonstrated validity and reliability, is the Creighton Competency Evaluation Instrument (C-CEI; Todd, Manz, Hawkins, Parsons, & Hercinger, 2008). The C-CEI includes 22 nursing behaviors that can be observed and evaluated in a simulation. These behaviors are grouped into four categories:

- Assessment (e.g., collection of pertinent information about the patient and the environment)
- Communication (e.g., with the simulated patient and team members, documentation, responses to abnormal findings)
- Clinical judgment (e.g., interpretation of vital signs and laboratory findings, performance and evaluation of interventions)
- Patient safety (e.g., patient identifiers, medications, technical performance)

The C-CEI is available at <https://nursing.creighton.edu/academics/competency-evaluation-instrument>.

Even with a validated tool, however, evaluators using it may not interpret the behaviors similarly nor score them as intended. With high-stakes evaluation, the evaluators need a shared mental model about the performance expected in the simulation and to agree on specific behaviors that would represent successful performance of the competencies (Oermann, Kardong-Edgren, & Rizzolo, 2016b). In one study, evaluators had extensive training on using the C-CEI for observing and rating performance in a simulation for high-stakes evaluation (Kardong-Edgren, Oermann, Rizzolo, & Odom-Maryon, 2017). The training extended over a period of time and included refreshers. Nine of 11 raters developed a shared mental model for scoring and were

consistent in their ratings of student performance. However, two raters, even with this extensive training, were outliers and were inconsistent with other evaluators and in their own scoring (intrarater reliability). Findings emphasized the importance of training faculty members or whoever is rating performance in the simulation on the tool and behaviors that would indicate successful performance of the competencies, and ensuring that all evaluators are competent to judge performance.

Tools for high-stakes evaluation with simulation can provide for analytic or holistic scoring similar to scoring for essay items and written assignments, which was discussed in earlier chapters. With analytic scoring, the evaluator observes student performance and typically rates each component of the performance. An example of analytic scoring is use of a skills checklist: The evaluator observes each step of the skill, verifying that it was performed correctly. Holistic scoring, in contrast, allows the evaluator to observe multiple behaviors in a simulation and rate the performance as a whole. Rating scales are examples of holistic scoring—these tools provide a means of rating a range of competencies, including some that are complex. The C-CEI is an example of a holistic tool used for assessing performance in a simulation. For some objectives, knowledge, and competencies to be assessed, multiple tools would be appropriate, some to rate skills in a simulation and others to provide a global rating of performance (Oermann et al., 2016a).

If the assessment is for high-stakes decisions, more than one evaluator should observe and rate performance. As discussed in earlier chapters, teachers may focus on different aspects of a performance. With more than one evaluator, the ratings can be combined, providing a fairer assessment for the student. Assessment for high-stakes decisions should be done by teachers who have not worked previously with the students. This avoids the chance of bias, positive or negative, when observing the performance. If the performance is video recorded, the evaluators can rate the performance independently to avoid influencing each other's scores.

Reliability is critical for a high-stakes assessment. With interrater reliability, different evaluators are consistent in their ratings of the performance and decisions about whether students are competent. There also should be intrarater reliability—if the evaluators observed the student a second time, those ratings would be similar to the first observation. It is generally easier to obtain reliability with an assessment of skills and procedures using a checklist than with a rating scale. The competencies on a rating scale are broader, for example, communicates effectively with providers, which allow for different interpretations and judgments (Oermann et al., 2016a).

Evaluators must be trained in assessment and use of the tool. This is critical to establish reliability. Everyone involved in the assessment needs to be aware of the objectives and the knowledge and competencies to be evaluated, and they need to have a shared understanding of the meaning of each item on the tool and what competent performance would “look like.” The observations and interpretations of performance by the evaluators must be accurate. Errors that can occur with rating scales

were presented in Chapter 14, Clinical Evaluation Methods. These also relate to simulation-based assessments. One type of error occurs when the evaluator only rates performance at the midpoint of the rating scale; this is an error of central tendency. Or evaluators may be too lenient, rating student performance in the simulation at the high end of the scale, or too severe, rating it at the low end, regardless of the quality of the performance. If the rating form has too many specific competencies on it or a checklist has too many discrete steps, a logical error might occur; the same rating is given for related items on the tool without the evaluator observing each one. Similar to clinical evaluation and grading essay items and written assignments, if evaluators know the student being observed in the simulation, they may have an impression of the student that influences their current evaluation of the performance (a halo effect). This is why one of the recommendations for high-stakes evaluations using simulation is that evaluators should not know the student being observed.

During the training, evaluators should discuss the tool and come to consensus about the meaning of each competency and performance expected in the simulation. This is critical to have a shared interpretation (a mental model) of what competent performance would look like. Evaluators should practice using the tool for rating performance, for example, rating the performance of a student in a video recording or on YouTube, and discuss their ratings with each other. When one evaluator is used for the assessment, the evaluator should practice using the tool and discuss ratings with colleagues to ensure similar interpretations of the competencies and observations. Exhibit 15.1 provides a summary of these key steps in using simulation for assessment.

EXHIBIT 15.1

DEVELOPING ASSESSMENTS USING SIMULATION

1. Identify the objectives of the assessment and the knowledge and skills to be evaluated.
2. Develop a simulation that requires use of that knowledge and those competencies.
3. Have colleagues and other experts review the simulation (scenario) to confirm that students will use the intended knowledge and competencies in it.
4. Select a tool (or tools) that produces reliable results that form the basis for sound decisions.
5. Train the evaluators to ensure their agreement on the meaning of the competencies and behaviors that would represent competent performance, and to make them aware of errors that can occur with ratings of performance.
6. Provide practice for evaluators to be comfortable with the tool and observation of performance in a simulation.
7. For high-stakes evaluation, arrange for evaluators who do not know the students being assessed.

■ Standardized Patients

One type of simulation for assessment of competencies uses standardized patients. Standardized patients are individuals who have been trained to accurately portray the role of a patient with a specific diagnosis or condition. With simulations using standardized patients, students can be assessed on a history and physical examination, related skills and procedures, and communication abilities, among other outcomes. Standardized patients are effective for evaluation because the actors are trained to recreate the same patient condition and clinical situation each time they are with a student, thereby providing consistency in the performance evaluation.

When standardized patients are used for formative assessment, they provide feedback to the students on their performance, an important aid to their learning. With the transition to distance education in nursing, technology is being used increasingly to allow students to interact with standardized patients when they are not on site (Ainslie & Bragdon, 2018; Ballman, Garritano, & Beery, 2016; Carman et al., 2017). Because they are trained for their role, standardized patients are well suited for summative assessment of students' clinical, interpersonal, and communication skills.

■ Objective Structured Clinical Examination

An OSCE provides a means of evaluating performance in a laboratory setting or via technology when students are not on site in the school, rather than in the clinical setting. OSCEs are an objective means of assessing performance of varied skills and competencies (Goh, Zhang, Lee, Wu, & Wang, 2018; Selim & Dawood, 2015). In an OSCE, students rotate through a series of stations; at each station, they complete an activity or perform a task, which is then evaluated. Some stations assess the student's ability to take a patient's history, perform a physical examination, and implement other interventions while being observed by the teacher or examiner (evaluator). The student's performance then can be rated using a checklist or rating scale. Checklists are the most widely used tool in an OSCE, followed by rating scales or a combination of these (Goh et al., 2018). At other stations, students might be tested on their knowledge and cognitive skills—they might be asked to analyze data, select interventions and treatments, and manage the patient's condition. Most often OSCEs are used for summative assessment; however, they also can be used formatively to assess performance and provide feedback to students. Meskell et al. (2015) indicated that OSCEs facilitate the assessment of psychomotor skills, knowledge, and attitudes of students. These authors also suggest that OSCE assessments help develop students' confidence in their performance and clinical skills and prepare them for clinical practice.

Different types of stations can be used in an OSCE. At one type of station, the student may interact with a standardized patient to collect a patient history and conduct a physical examination. At these stations, the teacher or examiner can evaluate students' understanding of varied patient conditions and management of them and can rate the students' performance. At other stations, students may demonstrate skills, perform procedures, use technologies, and demonstrate other technical competencies. Performance at these stations may be evaluated by the teacher or examiner, or a standardized patient. Lynga, Masiello, Karlgren, and Joelsson-Alm (2019) compared teacher and peer assessments in an OSCE of clinical skills among undergraduate nursing students. The OSCE protocol was a structured checklist to facilitate the assessment. There was a 94% agreement between these assessments using this checklist.

OSCEs are used frequently in nurse practitioner programs for competency assessment. In British Columbia, they are part of the licensure procedure for nurse practitioners. The OSCE uses 16 stations (15 stations for the adult and pediatric examination) to assess the knowledge, skills, and abilities of nurse practitioners to provide comprehensive and safe care at the entry level. The examination includes two types of stations: (a) couplet stations, where the nurse practitioners have a 5-minute encounter with a standardized patient followed by 5 minutes of a written follow-up related to the encounter, and (b) 10-minute stations, where they have a 10-minute encounter with a standardized patient (interactive stations) but no written component (College of Registered Nurses of British Columbia, 2018). The clinical skills assessed in the OSCE include history taking, patient education and counseling, clinical problem-solving, developing plans of care, documentation, physical examination, ordering and interpreting diagnostic tests, diagnosing, prescribing, and consulting and referral (p. 4).

There also may be post encounter stations to facilitate the evaluation of cognitive skills such as interpreting lab results and other data, developing management plans, and making other types of decisions about patient care. Students may be asked to document their findings with the standardized patient, answer questions about the clinical situation, and provide evidence for their decisions, among other competencies (Hawkins & Boulet, 2008). At these stations, the teacher or examiner is not present to observe students.

Mitchell et al. (2015) built on their earlier work to explore the use of OSCEs in undergraduate nursing education and to develop best practice guidelines for OSCEs. They applied these guidelines to modify OSCEs in a course and gather feedback from nursing students ($n = 691$) and lecturers ($n = 14$). The findings of the study supported the use of these guidelines for developing OSCEs in nursing education. Table 15.1 provides a list of the best practices that educators can use when developing OSCEs for undergraduate or graduate nursing students and for assessment of nurses' skills and competencies.

TABLE 15.1 Best Practice Guidelines for OSCEs

1. OSCEs should focus on common practices, attitudes, and skills (or significant encounters).
2. The knowledge, attitudes, and skills should be relevant to OSCE learning and assessment.
3. OSCEs should be structured and delivered in a manner that aligns directly with mastery of desired knowledge, attitudes, and skills.
4. OSCEs should be timed appropriately in the sequence of students' learning (to maximize synthesis of disparate course content and minimize the potential for a piecemeal, superficial learning approach).
5. OSCEs should be judged using a holistic rating tool.
6. Students in the OSCE should perform tasks in an integrated rather than piecemeal fashion (combining assessments of discrete skills in an authentic way).
7. The knowledge, attitudes, and skills should relate to the delivery of safe patient-centered care.
8. Educators should provide for ongoing practice of integrated assessment and intervention skills in a safe and supportive environment, with feedback to guide students' development and reflection.

OSCE, objective structured clinical examination.

Source: From Mitchell, M. L., Henderson, A., Jeffrey, C., Nulty, D., Groves, M., Kelly, M., ... Glover, P. (2015). Application of best practice guidelines for OSCEs—An Australian evaluation of their feasibility and value. *Nurse Education Today*, 6, 701. Reprinted by permission.

Students need to be prepared for an OSCE. This preparation is important to familiarize them with the process and reduce their stress during the assessment. Students should be clear about the knowledge and competencies to be assessed; should be familiar with the technology, equipment, checklists, and rating forms that might be used; and should be told about the number of stations and timing of the OSCE.

■ Summary

Simulations provide learning activities for students without the constraints of a real-life situation. With simulations, students can develop their psychomotor and technological skills and practice those skills to maintain their competence. Simulations, particularly those involving high-fidelity simulation, enable students to think through clinical situations and make independent decisions. With these experiences, students can develop their clinical-judgment skills. Simulations also can be used for assessment. Students can demonstrate procedures and technologies, analyze scenarios, make decisions about problems and actions to take, carry out interventions, and evaluate their decisions. Students' knowledge and performance in a simulation can be assessed for feedback or for summative purposes, verifying their competencies.

In using simulation for assessment, the first task is to identify the objectives of the assessment and the knowledge and competencies to be evaluated. The simulation planned for the assessment should reflect the objectives and require that students use that knowledge and those skills. Colleagues and other experts can review the simulation to ensure that it is appropriate for the objectives and for the knowledge and competencies to be assessed, and that it represents a realistic clinical scenario. This review helps establish the validity of the simulation. For assessment, it is critical that the tool for evaluating performance produces reliable results and that evaluators are trained in its use. With training, evaluators come to agreement as to the meaning of the competencies on the tool and behaviors that would indicate competent performance.

One type of simulation for assessment uses standardized patients, that is, individuals who have been trained to accurately portray the role of a patient with a specific diagnosis or condition. Another type of assessment of clinical competencies is an OSCE, in which students rotate through a series of stations completing activities or performing skills that are then evaluated. Guidelines for developing simulations for high-stakes evaluations and for using OSCEs were provided in the chapter.

■ References

- Ainslie, M., & Bragdon, C. (2018). Telemedicine simulation in online family nurse practitioner education: Clinical competency and technology integration. *Journal of the American Association of Nurse Practitioners*, 30, 430–434. doi:10.1097/jxx.0000000000000071
- Ballman, K., Garritano, N., & Beery, T. (2016). Broadening the reach of standardized patients in nurse practitioner education to include the distance learner. *Nurse Educator*, 41, 230–233. doi:10.1097/NNE.0000000000000260
- Boulet, J. R., & Murray, D. J. (2010). Simulation-based assessment in anesthesiology: Requirements for practical implementation. *Anesthesiology*, 112, 1041–1052. doi:10.1097/ALN.0b013e3181cea265
- Carman, M., Xu, S., Rushton, S., Smallheer, B. A., Williams, D., Amarasekara, S., & Oermann, M. H. (2017). Use of a virtual learning platform for distance-based simulation in an acute care nurse practitioner curriculum. *Dimensions of Critical Care Nursing*, 36, 284–289. doi:10.1097/dcc.0000000000000259
- Chmil, J. V., Turk, M., Adamson, K., & Larew, C. (2015). Effects of an experiential learning simulation design on clinical nursing judgment development. *Nurse Educator*, 40, 228–232. doi:10.1097/NNE.0000000000000159
- College of Registered Nurses of British Columbia. (2018). *Objective structured clinical examination (OSCE) candidate guidebook (family, adult, pediatric)*. Vancouver, BC: College of Registered Nurses of British Columbia. Retrieved from https://www.bccnp.ca/Registration/RN_NP/Documents/NPOSCECandidateGuidebook.pdf
- Dube, M. M., Reid, J., Kaba, A., Cheng, A., Eppich, W., Grant, V., & Stone, K. (2019). PEARLS for systems integration: A modified PEARLS framework for debriefing systems-focused simulations. *Simulation in Healthcare* 14, 333–342. doi:10.1097/sih.0000000000000381

- Feather, R. A., Carr, D. E., Reising, D. L., & Garletts, D. M. (2016). Team-based learning for nursing and medical students: Focus group results from an interprofessional education project. *Nurse Educator*, 41(4), E1–E5. doi:10.1097/NNE.0000000000000240
- Fey, M. K., & Jenkins, L. S. (2015). Debriefing practices in nursing education programs: Results from a national study. *Nursing Education Perspectives*, 36, 361–366.
- Furseth, P. A., Taylor, B., & Kim, S. C. (2016). Impact of interprofessional education among nursing and paramedic students. *Nurse Educator*, 41, 75–79. doi:10.1097/NNE.0000000000000219
- Gayle, D. (2019). In-simulation debriefing increases therapeutic communication skills. *Nurse Educator*. doi:10.1097/nne.0000000000000643 [e-pub ahead of print]
- Goh, H. S., Zhang, H., Lee, C. N., Wu, X. V., & Wang, W. (2018). Value of nursing objective structured clinical examinations: A scoping review. *Nurse Educator* 44, E1–E6. doi:10.1097/nne.0000000000000620
- Hawkins, R. E., & Boulet, J. R. (2008). Direct observation: Standardized patients. In E. S. Holmboe & R. E. Hawkins (Eds.), *Practical guide to the evaluation of clinical competencies* (pp. 102–118). Philadelphia, PA: Mosby.
- Hayden, J. K., Smiley, R. A., Alexander, M., Kardong-Edgren, S., & Jeffries, P. R. (2014). The NCSBN National Simulation Study: A longitudinal, randomized, controlled study replacing clinical hours with simulation in prelicensure nursing education. *Journal of Nursing Regulation*, 5(2), S3–S64.
- Horsley, T. L., Reed, T., Muccino, K., Quinones, D., Siddall, V. J., & McCarthy, J. (2016). Developing a foundation for interprofessional education within nursing and medical curricula. *Nurse Educator*, 41, 234–238. doi:10.1097/NNE.0000000000000255
- Kardong-Edgren, S. K., Oermann, M. H., & Rizzolo, M. A. (2019). Emerging theories influencing the teaching of clinical nursing skills. *Journal of Continuing Education in Nursing*, 50, 257–262. doi:10.3928/00220124-20190516-05
- Kardong-Edgren, S. K., Oermann, M. H., Rizzolo, M. A., & Odom-Maryon, T. (2017). Establishing inter- and intrarater reliability for high-stakes testing using simulation. *Nursing Education Perspectives*, 38, 63–68. doi:10.1097/01.Nep.0000000000000114
- Klenke-Borgmann, L. (2019). High-fidelity simulation in the classroom for clinical judgment development in third-year baccalaureate nursing students. *Nursing Education Perspectives*. doi:10.1097/01.Nep.0000000000000457. [e-pub ahead of print]
- Lasater, K. (2007). High-fidelity simulation and the development of clinical judgment: Students' experiences. *Journal of Nursing Education*, 46, 269–276.
- Lasater, K. (2011). Clinical judgment: The last frontier for evaluation. *Nurse Education in Practice*, 11, 86–92. doi:10.1016/j.nepr.2010.11.013
- Lee, N. J., Jang, H., & Park, S. Y. (2016). Patient safety education and baccalaureate nursing students' patient safety competency: A cross-sectional study. *Nursing & Health Sciences*, 18, 163–171. doi:10.1111/nhs.12237
- Lynga, P., Masiello, I., Karlgren, K., & Joelsson-Alm, E. (2019). Experiences of using an OSCE protocol in clinical examinations of nursing students—A comparison of student and faculty assessments. *Nurse Education in Practice*, 35, 130–134. doi:10.1016/j.nepr.2019.02.004
- Meskill, P., Burke, E., Kropmans, T. J., Byrne, E., Setyonugroho, W., & Kennedy, K. M. (2015). Back to the future: An online OSCE Management Information System for nursing OSCEs. *Nurse Education Today*, 35, 1091–1096. doi:10.1016/j.nedt.2015.06.010
- Mitchell, J. D., Amir, R., Montealegre-Gallegos, M., Mahmood, F., Shnider, M., Mashari, A., ... Matyal, R. (2018). Summative objective structured clinical examination assessment at the end of anesthesia residency for perioperative ultrasound. *Anesthesia and Analgesia*, 126, 2065–2068. doi:10.1213/ane.0000000000002826

- Mitchell, M. L., Henderson, A., Jeffrey, C., Nulty, D., Groves, M., Kelly, M., ... Glover, P. (2015). Application of best practice guidelines for OSCEs—An Australian evaluation of their feasibility and value. *Nurse Education Today*, 35, 700–705. doi:10.1016/j.nedt.2015.01.007
- Mudumbai, S. C., Gaba, D. M., Boulet, J. R., Howard, S. K., & Davies, M. F. (2012). External validation of simulation-based assessments with other performance measures of third-year anesthesiology residents. *Simulation in Healthcare*, 7, 73–80. doi:10.1097/SIH.0b013e31823d018a
- Oermann, M. H., Kardong-Edgren, S. K., & Rizzolo, M. A. (2016a). Summative simulated-based assessment in nursing. *Journal of Nursing Education*, 55(6), 323–328. doi:10.3928/01484834-20160516-04
- Oermann, M. H., Kardong-Edgren, S. K., & Rizzolo, M. A. (2016b). Towards an evidence-based methodology for high stakes evaluation of nursing students' clinical performance. *Teaching & Learning in Nursing*, 11, 133–137. doi:10.1016/j.teln.2016.04.001
- Oermann, M. H., Molloy, M., & Vaughn, J. (2014). Use of deliberate practice in teaching in nursing. *Nurse Education Today*, 35, 535–536. doi:10.1016/j.nedt.2014.11.007
- Oermann, M. H., Muckler, V. C., & Morgan, B. (2016). Framework for teaching psychomotor and procedural skills in nursing. *Journal of Continuing Education in Nursing*, 47(6), 278–282. doi:10.3928/00220124-20160518-10
- O'Leary, F. (2015). Simulation as a high stakes assessment tool in emergency medicine. *Emergency Medicine Australasia*, 27, 173–175. doi:10.1111/1742-6723.12370
- Owen, M. I., Garbett, M., Coburn, C. V., & Amar, A. F. (2017). Implementation of deliberate practice as a simulation strategy in nursing education. *Nurse Educator*, 42, 273–274. doi:10.1097/nne.0000000000000371
- Page-Cuttrara, K., & Turk, M. (2017). Impact of prebriefing on competency performance, clinical judgment and experience in simulation: An experimental study. *Nurse Education Today*, 48, 78–83. doi:10.1016/j.nedt.2016.09.012
- Reed, T., Horsley, T. L., Muccino, K., Quinones, D., Siddall, V. J., McCarthy, J., & Adams, W. (2016). Simulation using TeamSTEPPS to promote interprofessional education and collaborative practice. *Nurse Educator*, 42, E1–E5. doi:10.1097/nne.0000000000000350
- Reed, T., Pirotte, M., McHugh, M., Oh, L., Lovett, S., Hoyt, A. E., ... McGaghie, W. C. (2016). Simulation-based mastery learning improves medical student performance and retention of core clinical skills. *Simulation in Healthcare*, 11, 173–180. doi:10.1097/SIH.0000000000000154
- Rutherford-Hemming, T., & Lioce, L. (2018). State of interprofessional education in nursing: A systematic review. *Nurse Educator*, 43, 9–13. doi:10.1097/nne.0000000000000405
- Selim, A. A., & Dawood, E. (2015). Objective structured video examination in psychiatric and mental health nursing: A learning and assessment method. *Journal of Nursing Education*, 54, 87–95. doi:10.3928/01484834-20150120-04
- Sullivan, N. (2015). An integrative review: Instructional strategies to improve nurses' retention of cardiopulmonary resuscitation priorities. *International Journal of Nursing Education Scholarship*, 12. doi:10.1515/ijnes-2014-0012
- Tavares, W., Brydges, R., Myre, P., Prpic, J., Turner, L., Yelle, R., & Huiskamp, M. (2018). Applying Kane's validity framework to a simulation based assessment of clinical competence. *Advances in Health Sciences Education*, 23, 323–338. doi:10.1007/s10459-017-9800-3
- Todd, M., Manz, J., Hawkins, K., Parsons, M., & Hercinger, M. (2008). The development of a quantitative evaluation tool for simulations in nursing education. *International Journal of Nursing Education Scholarship*, 5, Article 41. doi:10.2202/1548-923X.1705
- Victor, J. (2017). Improving clinical nursing judgment in prelicensure students. *Journal of Nursing Education*, 56, 733–736. doi:10.3928/01484834-20171120-05

V

ISSUES RELATED TO TESTING AND EVALUATION IN NURSING EDUCATION

SOCIAL, ETHICAL, AND LEGAL ISSUES

Over the past decade, educational testing and assessment have grown in use and importance for students in general, and nursing students in particular. One only has to read the newspapers and watch television to appreciate the prevalence of testing and assessment in contemporary American society. With policies and laws such as the No Child Left Behind Act, mandatory high school graduation tests in some states, and the emphasis on standardized achievement tests in many schools, testing and assessment have taken a prominent role in the educational system. From the moment of birth, when we are weighed, measured, and rated according to the Apgar scale, throughout all of our educational and work experiences, and even in our personal and social lives, we are used to being tested and evaluated. In addition, nursing and other professional disciplines have come under increasing public pressure to be accountable for the quality of educational programs and the competency of their practitioners; thus, testing and assessment often are used to provide evidence of quality and competence.

With the increasing use of assessment and testing come intensified interest and concern about fairness, appropriateness, and impact. This chapter discusses selected social, ethical, and legal issues related to testing and assessment practices in nursing education.

■ Social Issues

Testing has tremendous social impact because test scores can have positive and negative consequences for individuals. Tests can provide information to assist in decision-making; some of these decisions have more importance to society and to individuals than other decisions. The licensure of drivers is a good example. Written and performance tests provide information for deciding who may drive a vehicle. Society has a vested interest in the outcome because a bad decision can affect the safety of a great many people. Licensure to drive a vehicle also may be an important issue to an individual; some jobs require the employee to drive a car or truck, so a person who lacks a valid operator's license will not have access to these employment opportunities.

Tests also are used to help to place individuals into professional and occupational roles. These placement decisions have important implications because a person's profession or occupation to some extent determines status and economic and political power. Because modern society depends heavily on scientific knowledge and technical competence, occupational and professional role selection are based to a significant degree on what individuals know and can do. Therefore, by controlling who enters certain educational programs, institutions have a role in determining the possible career path of an individual.

The way in which schools should select candidates for occupational and professional roles is a matter of controversy, however. Some individuals and groups hold the view that schools should provide equal opportunity and access to educational programs. Others believe that equal opportunity is not sufficient to allow some groups of people to overcome discrimination and oppression that has handicapped their ability and opportunity.

Decisions about which individuals should be admitted to a nursing education program are important because of the nursing profession's commitment to the good of society and to the health and welfare of current and future patients (American Nurses Association, 2010). Nursing faculties must select individuals for admission to nursing programs who are likely to practice nursing competently and safely; tests frequently are used to assist educators in selecting candidates for admission. Improper use of testing or the misinterpretation of test scores can result in two types of poor admission decisions. If an individual is selected who is later found to be incompetent to practice nursing safely, the public might be at risk; if an individual who would be competent to practice nursing is not admitted, that individual is denied access to a professional role.

The use of testing in employment situations and for the purpose of professional certification can produce similar results. Employers have a stake in making these decisions because they are responsible for ensuring the competence of their employees. Tests for employment, to ensure competencies at the end of orientation, and to certify continuing knowledge and skills are important not only to the employee but also to the employer. Through assessments such as these, the employer verifies that the individual is competent for the role. Selection decisions, therefore, have social implications for individuals, institutions, and society as a whole.

Although educational and occupational uses of testing are growing in frequency and importance, the public often expresses concerns about testing. Some of these concerns are rational and relevant; others are unjustified.

Assessment Bias

One common concern is that assessments are biased or unfair to certain groups of test-takers. A major purpose of assessment is to discriminate among people, that

is, to identify important differences among them with regard to their knowledge, skills, or attitudes. To the extent that differences in scores represent real differences in achievement of objectives, this discrimination is not necessarily unfair. Bias can occur, however, when scores from an assessment are misinterpreted, or conclusions are drawn about performance that go well beyond the assessment. For example, if a test is found to discriminate between men and women on variables that are not relevant to educational or occupational success, it would be unfair to use that test to select applicants for admission to a program or for a job. Thus, the question of test bias really is one of measurement validity, the degree to which inferences about test results are justifiable in relation to the purpose and intended use of the test (Brookhart & Nitko, 2019; Miller, Linn, & Gronlund, 2013).

Assessment bias also has been defined as the differential validity of an assessment result for a group of students or other people being assessed. With assessment bias, a given score does not have the same meaning for all students who were assessed. The teacher may interpret a low test score to mean inadequate knowledge of the content, but there may be a relevant subgroup of individuals, for example, students with learning disabilities, for whom that score interpretation is not accurate. The test score may be low for a student with a learning disability because he or she did not have enough time to complete the exam or because there was too much environmental noise, not because of a lack of knowledge about the content.

Individual test items also can discriminate against subgroups of test-takers, such as students from ethnic minority groups; this is termed *differential item functioning* (DIF). Test items are considered to function differentially when students of different subgroups but of equal ability, as evidenced by equal total test scores, perform differently on the item. However, differences in item functioning do not necessarily confirm item bias (Brookhart & Nitko, 2019). Although some experts suggest excluding DIF items from a test, eliminating such items may decrease evidence of content and construct validity, especially if there is a large number of DIF items (Kabasacal & Kelecioğlu, 2015).

Item bias (and collectively, test bias) also can be construed as a content and experience differential. Bias is produced by test or item content that differs substantially from one subgroup's life experiences *and* when these differences are not taken into account when the assessment results are interpreted (Brookhart & Nitko, 2019). The presence of test bias does not necessarily mean that the test is unfair. For a test to be unfair, the bias among test scores for different groups occurs when (a) scores are not used and interpreted the same across all students, (b) the opportunity to prepare for or complete the test is not the same for all students, and (c) the test conditions are not uniform for all students (Balkin, Heard, Lee, & Wines, 2014).

A culturally biased item contains references to a particular culture and is more likely to be answered incorrectly by students from a minority group. An example of a culturally biased test item follows:

1. While discussing her health patterns with the nurse, a patient says that she enjoys all of the following leisure activities. Which one is an aerobic activity?
 - a. Attending ballet performances
 - b. Cultivating house plants
 - c. Line dancing
 - d. Singing in the church choir

The correct answer is “line dancing,” but students who are non-native English speakers or English-language learners, students from cultural minority groups, and even domestic students from certain regions of the country may be unfamiliar with this term and therefore may not select this response. In this case, an incorrect response may mean that the student is unfamiliar with this type of dancing, not that the student is unable to differentiate between aerobic and nonaerobic activities. As discussed in Chapter 2, Qualities of Effective Assessment Procedures: Validity, Reliability, and Usability, cultural bias of this type contributes to construct-irrelevant variance that can reduce measurement validity (Miller et al., 2013). The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [NCME], 2014) specify that test developers should reduce language differences that threaten the reliability and validity of inferences made from test scores.

Careful peer review of test items for discernible bias allows the teacher to reword items to remove references to American or English literature, music, art, history, customs, or regional terminology that are not essential to the nursing content being tested. The inclusion of jokes, puns, and other forms of humor also may contribute to cultural bias because these forms of expression may not be interpreted correctly by international students, non-native English speakers, and English-language learners. It is appropriate, however, to include content related to cultural differences that are essential to safe nursing practice. Students and graduate nurses must be culturally competent if they are to meet the needs of patients from a variety of cultures.

A test item with linguistic or structural bias is poorly written. It may be lengthy, unclear, or awkwardly worded, interfering with the student’s understanding of the teacher’s intent. Structurally biased items create problems for all students, but they are more likely to discriminate against English-language learners or those with learning disabilities. In addition, students from minority cultures may be less likely than dominant-culture students to ask the test proctor to clarify a poorly written item,

usually because it is inappropriate to question a teacher in certain cultures. Following the general rules for writing test items in this book will help the teacher to avoid structural bias.

An assessment practice that helps to protect students from potential bias is anonymous or blinded scoring and grading. The importance of scoring essay items and written assignments anonymously was discussed earlier in the book. Anonymous grading also can be used for an entire course. The process is similar to that of peer review of manuscripts and grant proposals: The teacher is unaware of the student's identity until the end of the course. Students choose a number or are randomly assigned an anonymous grading system number at the beginning of a course. That number is recorded on every test, quiz, written assignment, and other assessments during the semester, and scores are recorded according to these code numbers. The teacher does not know the identity of the students until the end of the course. This method of grading prevents the influence of a teacher's previous impressions of a student on the scoring of a test or written assignment.

Grade and Test Score Inflation

Another common criticism of testing concerns the general trend toward inflation of test scores and grades at all educational levels. Grade inflation is the tendency to award higher grades over time for performance that does not improve in quality. Another sense of this term is *grading leniency*, which is the tendency to award higher grades to students than they deserve. Grade inflation distorts the meaning of test scores, making it difficult for teachers to use them wisely in decision-making. If an A is intended to represent exceptional or superior performance, then all students cannot earn A grades because if everyone is exceptional, then no one is. With grade inflation, all grades are compressed near the top, which makes it difficult to discriminate among students. When there is little distribution of scores or grades, there is little value in testing. Issues common to the problem of grade inflation include:

- Students' expectations related to the belief that they are consumers of the educational program (Edgar, Johnson, Graham, & Dixon, 2014; Nata, Pereira, & Neves, 2014) and the desire to have a greater role in their own educations (Holman, 2015)
- The growing importance of grades for graduate school and employment applications (Holman, 2015)
- Mandatory faculty evaluation and the threat of negative student evaluations; faculty members "wanting to be liked" (Docherty, 2018; Holman, 2015)
- Faculty aversion to giving failing grades (Couper, 2018; Paskausky & Simonelli, 2014)

- Faculty beliefs about what constitutes satisfactory performance (the “good enough” approach) and the subjective nature of grading (Docherty, 2018)
- Faculty anxiety about students’ seeking legal recourse for a failing grade (Couper, 2018)
- The “hassle factor” associated with assigning a failing grade (Docherty, 2018)
- Faculty perception of absence of administrative support for assigning failing grades (Couper, 2018; Hughes, Mitchell, & Johnston, 2016)
- Differences in grading practices of tenured and nontenured faculty members (Donaldson & Gray, 2012; Paskausky & Simonelli, 2014)
- Increasing use of part-time faculty members in nursing education programs

The relationship among these factors is especially relevant in nursing education. Most part-time faculty members teach in the clinical area, and many are skilled clinicians with little or no formal academic preparation for the role of educator. Many nursing faculty members are reluctant to assign failing grades in clinical courses, giving students the benefit of the doubt especially in beginning courses. This belief is easily communicated to part-time faculty members, who may have additional concerns about their job security because most of them are hired on limited-term contracts. Where student evaluation of faculty members is mandatory, part-time teachers may be unwilling to assign lower clinical grades because of possible repercussions related to continued employment in that role.

In addition, grading discrepancies between theory and related clinical courses frequently occur. Especially in nursing education programs where clinical practice is assigned a letter grade (instead of a pass–fail or similar grading system), higher clinical grades tend to inflate the overall grade point average. This discrepancy is difficult to explain or defend on the basis of the assumption that theory informs clinical practice; why would a student with a grade of C in a theory course be likely to earn an A grade in the corresponding clinical course? Clinical grade inflation of this sort may result in more students with marginal ability “slipping through the cracks” and failing the final clinical course of the nursing education program, or graduating only to fail the National Council Licensure Examination (NCLEX®).

In a study of discrepancy between the scores on a final exam and the clinical grades assigned by faculty members in an undergraduate program, Paskausky and Simonelli (2014) found a moderate to low correlation between the two measurements. The clinical grade distribution was negatively skewed with a narrow range; the test scores were normally distributed with a wide range. Grade discrepancy scores were calculated by subtracting the final exam score from the clinical grade; 98% of students had clinical grades higher than final exam grades.

Clinical grading also may be governed by the “rule of C,” where the D grade is virtually eliminated as a grading option because of program policies that require

a minimum grade of C to pass a clinical course. As previously mentioned, faculty members who are reluctant to assign failing grades to students then may award C grades to students with marginal performance, and the B grade becomes the symbol for average or acceptable performance. This grade compression (only three grade levels instead of five) contributes to grade inflation (Donaldson & Gray, 2012).

Another factor contributing to grade inflation is the increasing pressure of accountability for educational outcomes. When the effectiveness of a teacher's instruction is judged on the basis of students' test performance, the teacher may "teach to the test." Teaching to the test may involve using actual test items as practice exercises, distributing copies of a previously used test for review and then using the same test, or focusing exclusively on test content in teaching.

Because regulatory and accreditation standards for nursing education programs commonly include expectations of an acceptable first-time NCLEX pass rate for graduates each year, and the quality of graduate nursing programs is judged by graduates' pass rates on certification exams, these test results have significant implications for the educational institutions as well as the individual test-takers. When faculty members and educational programs are judged by how well their graduates perform on these high-stakes assessments, "direct preparation for the tests and assessments is likely to enter into classroom activities and thereby distort the curriculum" (Miller et al., 2013, p. 15).

It is important, however, to distinguish between teaching to the test and purposeful teaching of content to be sampled by the test and the practice of relevant test-taking skills. Nursing faculty members who understand the NCLEX test plan and ensure that their nursing curricula include content and learning activities that will enable students to be successful on the NCLEX are not teaching to the test.

Effect of Tests and Grades on Self-Esteem

Some critics of tests claim that testing results in emotional or psychological harm to students. The concern is that tests threaten students and make them anxious, fearful, and discouraged, resulting in harm to their self-esteem. There is no empirical evidence to support these claims. Feelings of anxiety about an upcoming test are both normal and helpful to the extent that they motivate students to prepare thoroughly so as to demonstrate their best performance. Because testing is a common life event, learning how to cope with these challenges is a necessary part of student development. Giving effusive praise for every performance whether or not it is praiseworthy (i.e., "ovation inflation"), while temporarily raising self-esteem, does little to produce students who can realistically assess their own efforts and persist despite challenges.

Brookhart and Nitko (2019) identified three types of test-anxious students: (a) students who have poor study skills and become anxious prior to a test because they do not understand the content that will be tested, (b) students who have good

study skills and understand the content but fear they will do poorly no matter how much they prepare for the exam, and (c) students who believe that they have good study skills but in essence do not. If teachers can identify why students are anxious about testing, they can direct them to specific resources such as those on study skills, test-taking strategies, and techniques to reduce their stress.

Most nursing students will benefit from developing good test-taking skills, particularly learners who are anxious. For example, students should be told to follow the directions carefully, read the item stems and questions without rushing to avoid misreading critical information, read each option for multiple-choice items before choosing one, manage time during the test, answer easy items first, and check their answers. Arranging the test with the easy items first often helps relieve anxiety as students begin the test. Because highly anxious students are easily distracted (Brookhart & Nitko, 2019), the teacher should ensure quiet during the testing session.

General guidelines for the teacher to follow to intervene with students who have test anxiety include (Brookhart & Nitko, 2019):

1. Identify the problem to be certain it is test anxiety and not a learning disability or a problem such as depression
2. Give specific detailed feedback about the student's performance on each test
3. Help the student to develop testwiseness skills (e.g., using time well, avoiding technical and clerical errors, learning how to make informed guesses, using unintended cues in item content or structure)
4. Refer the student to outside resources as needed
5. Advise students to concentrate on the assessment tasks and not allow themselves to be distracted (Brookhart & Nitko, 2019)

Although it is probably true that a certain level of self-esteem is necessary before a student will attempt the challenges associated with nursing education, high self-esteem is not essential to perform well on a test. In fact, if students are able to perform at their best, their self-esteem is enhanced. An important part of a teacher's role is to prepare students to do well on tests by helping them improve their study and test-taking skills and to learn to manage their anxiety.

Testing as a Means of Social Control

All societies sanction some form of social control of behavior; some teachers use the threat of tests and the implied threat of low test grades to control student behavior. In an attempt to motivate students to prepare for and attend class, a teacher may decide to give unannounced tests; the student who is absent that day will earn a score of zero, and the student who does not do the assigned readings will likely earn a low score. This practice is unfair to students because they need sufficient time to prepare

for a test to demonstrate their maximum performance, as discussed in Chapter 3, Planning for Testing. Students have a right to be informed in advance about when a test will be administered, and using tests in a punitive, threatening, or vindictive way is unethical (Brookhart & Nitko, 2019).

■ Ethical Issues

Ethical standards make it possible for nurses and patients to achieve understanding of and respect for each other. These standards also should govern the relationships of teachers and students. Contemporary bioethical standards include those of autonomy, freedom, veracity, privacy, beneficence, nonmaleficence, and fidelity. Several of these standards are discussed here as they apply to common issues in testing and evaluation.

The standards of privacy, autonomy, and veracity relate to the ownership and security of tests and test results. Some of the questions that have been raised are: Who owns the test? Who owns the test results? Who has or should have access to the test results? Should test-takers have access to standardized test items and their own responses?

Because educational institutions and employers started using standardized tests to make decisions about admission and employment, the public has been concerned about the potential discriminatory use of test results. The result of this public concern was the passage of federal and state “truth in testing” laws, requiring greater access to tests and test results.

Test-takers have the right to expect that certain information about them will be held in confidence. Teachers, therefore, have an obligation to maintain a privacy standard regarding students’ test scores. Such practices as public posting of test scores and grades should be examined in light of this privacy standard. Teachers should not post assessment results if individual students’ identities can be linked with their results; for this reason, many educational programs do not allow scores to be posted with student names or identification numbers. During posttest discussions, teachers should not ask students to raise their hands to indicate whether they answered an item correctly or incorrectly; this practice can be considered an invasion of students’ privacy (Brookhart & Nitko, 2019).

An additional privacy concern relates to the practice of keeping student records that include test scores and other assessment results. Questions often arise about who should have access to these files and the information they contain. Access to a student’s test scores and other assessment results is limited by laws such as the Family Educational Rights and Privacy Act of 1974 (FERPA). This federal law gives students certain rights with respect to their education records. For example, they can review their education records maintained by the school and request that the school correct records they believe to be inaccurate or misleading. Schools must have written

permission from the student to release information from the student's record except in selected situations such as accreditation or for program assessment purposes (U.S. Department of Education, n.d.). The FERPA limits access to a student's records to those who have legitimate rights to the information to meet the educational needs of the student. This law also specifies that a student's assessment results may not be transferred to another institution without written authorization from the student. In addition to these limits on access to student records, teachers should ensure that the information in the records is accurate and should correct errors when they are discovered. Files should be purged of anecdotal material when this information is no longer needed (Brookhart & Nitko, 2019).

Another way to violate students' privacy is to share confidential information about their assessment results with other teachers. To a certain extent, a teacher should communicate information about a student's strengths and weaknesses to other teachers to help them meet that student's learning needs. In most cases, however, this information can be communicated through student records to which other teachers have legitimate access. Informal conversations about students, especially if those conversations center on the teacher's impressions and judgments rather than on verifiable data such as test scores, can be construed as gossip.

Test results sometimes are used for research and program evaluation purposes. As long as students' identities are not revealed, their scores usually can be used for these purposes (Brookhart & Nitko, 2019). One way to ensure that this use of test results is ethical is to announce to the students when they enter an educational program that test results occasionally will be used to assess program effectiveness. Students may be asked for their informed consent for their scores to be used, or their consent may be implied by their voluntary participation in optional program evaluation activities. For example, if a questionnaire about student satisfaction with the program is distributed or mailed to students, those who wish to participate simply complete the questionnaire and return it; no written consent form is required. In many institutions of higher education, however, this use of test results may require review by the institutional review board.

The ethical principle of fidelity requires faithfulness in relationships and matters of trust. In nursing education programs, adherence to this principle requires that faculty members act in the best interest of students. By virtue of their education, experience, and academic position, faculty members hold power over their students. They have the ability to influence students' progress through the nursing education program and their ability to gain employment after graduation. Violations of professional boundaries may occur and affect students' ability to trust faculty members. Teachers who have personal relationships with students may be accused of awarding grades based on favoritism or, conversely, may be accused of using failing grades to retaliate against students who rebuff a sexual or emotional advance.

Standards for Ethical Testing Practice

Several codes of ethical conduct in using tests and other assessments have been published by professional associations, one of which is the *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 2004), reproduced in Appendix C. The *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014) describe standards for test construction, administration, scoring, and reporting; supporting documentation for tests; fairness in testing; and a range of testing applications. The *Standards* also address testing individuals with disabilities and different linguistic backgrounds. Common elements of these codes and standards are:

- Students have the right to be tested with tests that have been developed to meet professional standards.
- Teachers are responsible for the quality of the tests they develop and for selecting tests that are appropriate for the intended use.
- Test administration procedures must be fair to all students and protect their safety, health, and welfare.
- Teachers are responsible for the accurate scoring of tests and reporting test results to students in a timely manner.
- Students should receive prompt and meaningful feedback.
- Test results should be interpreted and used in valid ways.
- Teachers must communicate test results accurately and anticipate the consequences of using results to minimize negative results to students.
- Students must be able to present their concerns about the testing process or results and have those concerns reviewed seriously. (Brookhart & Nitko, 2019)

High-Stakes Assessments

High-stakes assessments are used for decision-making that results in serious consequences for the test-takers, teachers, and administrators of educational programs (Brookhart & Nitko, 2019). As mentioned in the introduction to this chapter, public demands for accountability have influenced a growing use of assessments intended to demonstrate that students and graduates meet knowledge and performance standards. In nursing as in many other health professions, an ongoing concern is the need to protect the public through the use of standardized licensure examinations to ensure competence to practice. State boards of nursing, which have regulatory authority to license nurses, share accountability with nursing education programs to

ensure the competence of program graduates. A state board of nursing's oversight of nursing education programs gives it the authority to curtail programs that have low licensure examination first-time pass rates. Therefore, licensure examinations are high-stakes assessments not only for nursing students and new graduates, but also for faculty members and administrators of nursing education programs (National Council of State Boards of Nursing [NCSBN®], 2007; National League for Nursing [NLN], 2010, 2012b).

Adding to the pressure on nursing education programs to meet state board of nursing minimum first-time pass rate requirements is a board of nursing's responsibility and authority also to sanction programs that have high attrition rates (NCSBN, 2007). The need to meet both standards (high NCLEX-RN® pass rate and low attrition rate) is a double-edged sword that has motivated increasing numbers of nursing education programs to use standardized end-of-program tests to identify students at risk of NCLEX-RN failure, and to deny program progression or graduation to students who are predicted to fail the licensure examination. According to Giddens and Morton (2010), this type of high-stakes assessment "borders on unethical educational practice" (p. 374) based on the need for multiple approaches to assessment of knowledge and skill when high-stakes decisions are based on the assessment (NLN, 2012b). In addition, use of a single standardized test as a basis for progression and graduation decisions raises concerns about whether any such test reliably predicts success among various subgroups of an increasingly diverse group of learners (Brookhart & Nitko, 2019).

Requiring students to achieve a predetermined score on a standardized test to graduate from the nursing education program or to be authorized to take the NCLEX so that the program's first-time pass rates meet or exceed a state board-mandated level is a complex problem for those who have successfully met all other program requirements. If this "exit exam" is a required component of a nursing course, students who cannot achieve the required score may fail the course, endangering their academic status. Students may need to take the exit examination repeatedly until they meet the standard, delaying graduation or licensure and thus adversely affecting them economically (NLN, 2012b). Also, most of the standardized tests being used as exit examinations are intended to predict whether an individual student is likely to *pass* the NCLEX. Such tests are much less accurate in predicting the likelihood of failure.

Progression or graduation policies requiring high-stakes testing also can distort the intended purpose of NCLEX pass-rate requirements as a measure of program quality. A nursing program that achieves a high first-time pass rate by allowing only the highest performing students to progress in the program, graduate, and take the NCLEX illustrates the effect of selection bias. Thus, the use of high-stakes assessments in progression and graduation policies raises concern about the extent to which the

nursing education program provides equal opportunity and access to diverse groups of students.

The NLN's concern about the "prevalent use of standardized tests to block graduation or in some other way deny eligibility to take the licensing exam" (NLN, 2012b) prompted creation of fair testing guidelines to assist nursing faculty members and administrators in developing and implementing ethical academic progression and graduation policies. These guidelines, reprinted in Appendix D, emphasize the obligation of faculty members and administrators to:

- Use multiple approaches for assessment of knowledge and clinical abilities when making high-stakes decisions.
- Select tests with evidence of measurement validity, and fairness and equity demonstrated by test performance across cultural, racial, or gender subgroups.
- Inform students about how the test results will be used.
- Undertake a comprehensive review of factors leading to development and implementation of high-stakes testing.
- Review other factors that affect NCLEX-RN pass rates and other measures of program quality, such as admissions policies, instructional effectiveness, remediation requirements, and course-level assessments, to identify opportunities for improvement. (NLN, 2012a)

■ Legal Aspects of Assessment

It is beyond the scope of this book to interpret laws that affect the use of tests and other assessments, and the authors are not qualified to give legal advice to teachers concerning their assessment practices. However, it is appropriate to discuss a few legal issues to provide guidance to teachers in using tests.

A number of issues have been raised in the courts by students claiming violations of their rights by testing programs. These issues include race or gender discrimination, violation of due process, unfairness of particular tests, various psychometric aspects such as measurement validity and reliability, and accommodations for students with disabilities (Brookhart & Nitko, 2019).

Psychometric Issues

Students seeking legal recourse for a failing grade in a course or on a high-stakes test may present concerns about technical or psychometric issues to the court. These issues may involve:

- Adherence to testing standards such as the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014) or the *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 2004)
- The measurement reliability and reliability of a test
- Test development procedures
- Methods of determining a passing score
- Issues with the physical features of the test, such as the quality of directions or print size
- Accommodations for students with disabilities (Brookhart & Nitko, 2019)

Assessment of Students With Disabilities

The Americans with Disabilities Act (ADA) of 1990 and the ADA Amendments Act of 2008 (P.L. 110–325) have influenced testing and assessment practices in nursing education and employment settings. This law prohibits discrimination against qualified individuals with disabilities. A *qualified individual with a disability* is defined as a person with a physical or mental impairment that substantially limits major life activities. Qualified individuals with disabilities meet the requirements for admission to and participation in a nursing program. Nursing education programs have a legal and an ethical obligation to accept and educate qualified individuals with disabilities. It is up to the nursing education program to provide reasonable accommodations, additional services and aids as needed, and to remove any barriers. This does not mean that institutions lower their standards to comply with the ADA.

The ADA requires teachers to make reasonable accommodations for disabled students to assess them properly. Such accommodations may include oral testing, computer testing, modified answer format, extended time for exams, test readers or sign language interpreters, a private testing area, or the use of large type for printed tests (Brookhart & Nitko, 2019; May, 2014). However, nursing faculty members should provide accommodations only if a student submits verification of qualification for such accommodations. This verification should be provided by the institutional officer responsible for disability services after receipt of evidence of the student's disability and individual needs from an appropriate professional. NCLEX policies permit test-takers with documented learning disabilities to have extended testing time as well as other reasonable accommodations, if approved by the board of nursing in the states in which they apply for initial licensure (NCSBN, 2019). This approval usually is granted only when the educational institution has verified the documentation of a disability and students' use of accommodations during the nursing education program. Because English-language proficiency is required for competent nursing

practice in the United States, non-native English speakers or English-language learners are not considered to be qualified persons with disabilities.

A number of concerns have been raised regarding the provision of reasonable testing accommodations for students with disabilities. One issue is the validity of the test result interpretations if the test was administered under standard conditions for one group of students and under accommodating conditions for other students. Any changes made to test items, administration conditions, or student response modes in order to accommodate students with disabilities must provide the teacher with valid assessment information (Brookhart & Nitko, 2019). The privacy rights of students with disabilities is another issue: Should the use of accommodating conditions be noted along with the student's test score? Such a notation would identify the student as disabled to anyone who had access to the record. There are no easy answers to such questions. In general, faculty members should be guided by accommodation policies developed by their institution and have any additional policies reviewed by legal counsel to ensure compliance with the ADA.

■ Summary

Educational testing and assessment are growing in use and importance for society in general and for nursing in particular. Nursing has come under increasing public pressure to be accountable for the quality of educational programs and the competency of its practitioners, and testing and assessment often are used to provide evidence of quality and competence. With the increasing use of assessment and testing come intensified interest in and concern about fairness, appropriateness, and impact.

The social impact of testing can have positive and negative consequences for individuals. Tests can provide information to assist in decision-making, such as selecting individuals for admission to education programs or for employment. The way in which selection decisions are made can be a matter of controversy, however, regarding equality of opportunity and access to educational programs and jobs.

The public often expresses concerns about testing. Common criticisms of tests include: tests are biased or unfair to some groups of test-takers; test scores have little meaning because of grade inflation; testing causes emotional or psychological harm to students; and tests are sometimes used in a punitive, threatening, or vindictive way. By understanding and applying codes for the responsible and ethical use of tests, teachers can ensure the proper use of assessment procedures and the valid interpretation of test results. Teachers must be responsible for the quality of the tests they develop and for selecting tests that are appropriate for their intended use. The use of high-stakes testing in progression and graduation policies is of particular concern, and guidelines are available to assist faculty members to develop fair testing policies.

The ADA of 1990 and the ADA Amendments Act of 2008 have implications for the proper assessment of students with physical and mental disabilities. This law requires educational programs to make reasonable testing accommodations for qualified individuals with learning as well as physical disabilities.

■ References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- American Nurses Association. (2010). *Nursing's social policy statement: The essence of the profession* (3rd ed.). Washington, DC: Author.
- Balkin, R. S., Heard, C. C. C., Lee, S., & Wines, L. A. (2014). A primer for evaluating test bias and test fairness: Implications for multicultural assessment. *Journal of Professional Counseling, Practice, Theory, & Research*, 4, 42–52.
- Brookhart, S. M., & Nitko, A. J. (2019). *Educational assessment of students* (8th ed.). New York, NY: Pearson Education.
- Couper, J. (2018). The struggle is real: Investigating the challenge of assigning a failing clinical grade. *Nursing Education Perspectives*, 39, 132–137.
- Docherty, A. (2018). Failing to fail in undergraduate nursing: Understanding the phenomenon. *Nursing Education Perspectives*, 39, 335–342.
- Donaldson, J. H., & Gray, M. (2012). Systematic review of grading practice: Is there evidence of grade inflation? *Nurse Education in Practice*, 12, 101–114.
- Edgar, L. D., Johnson, D. M., Graham, D. L., & Dixon, B. L. (2014). Student and faculty perceptions of plus/minus grading and its effect on course grade point averages. *College Student Journal*, 48(1), 184–197.
- Giddens, J. F., & Morton, N. (2010). Report card: An evaluation of a concept-based curriculum. *Nursing Education Perspectives*, 31, 372–377.
- Holman, D. (2015, March 12). Economic in-tuition grade inflation: False success destines us for real failure. *University Wire*. Retrieved from <http://search.proquest.com.authenticate.library.duq.edu/docview/1676444832?accountid=10610>
- Hughes, L. J., Mitchell, M., & Johnston, A. N. B. (2016, September). “Failure to fail” in nursing—A catch phrase or a real issue? A systematic integrative literature review. *Nurse Education in Practice*, 20, 54–63.
- Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. Washington, DC: American Psychological Association.
- Kabasacal, K. A., & Kelecioğlu, H. (2015). Effect of differential item functioning on test equating. *Educational Sciences: Theory & Practice*, 15, 1229–1246.
- May, K. A. (2014). Nursing faculty knowledge of the Americans with Disabilities Act. *Nurse Educator*, 39, 241–245.
- Miller, M. D., Linn, R. L., & Gronlund, N. E. (2013). *Measurement and assessment in teaching* (11th ed.). Upper Saddle River, NJ: Prentice Hall.
- Nata, G., Pereira, M. J., & Neves, T. (2014). Unfairness in access to higher education: A 11 year comparison of grade inflation by private and public secondary schools in Portugal. *Higher Education*, 68, 851–874.

- National Council of State Boards of Nursing. (2007). *Guiding principles of nursing regulation*. Retrieved from https://www.ncsbn.org/Guiding_Principles.pdf
- National Council of State Boards of Nursing. (2019). *2019 NCLEX® Examination candidate bulletin*. Retrieved from https://www.ncsbn.org/2019_Bulletin_Final.pdf
- National League for Nursing. (2010). *High-stakes testing*. Retrieved from <http://www.nln.org/advocacy-public-policy/issues/high-stakes-testing>
- National League for Nursing. (2012a). *NLN fair testing guidelines for nursing education*. Retrieved from <http://www.nln.org/docs/default-source/advocacy-public-policy/fair-testing-guidelines.pdf?sfvrsn=0>
- National League for Nursing. (2012b). *The fair testing imperative in nursing education*. Retrieved from http://www.nln.org/docs/default-source/about/nln-vision-series-%28position-statements%29/nlnvision_4.pdf
- Paskausky, A. L., & Simonelli, M. C. (2014). Measuring grade inflation: A clinical grade discrepancy score. *Nurse Education in Practice*, 14, 374–379.
- U.S. Department of Education. (n.d.). *Family Educational Rights and Privacy Act (FERPA)*. Washington, DC: Family Policy Compliance Office, U.S. Department of Education. Retrieved from <http://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html>

GRADING

The teacher's assessment of students provides the basis for assigning a grade for the course. The grade is a symbol that reflects the achievement of students in the course. In addition to grading the course as a whole, grades are given for individual assignments, quizzes, tests, and other learning activities completed by students throughout the course. This chapter examines the uses of grades in nursing programs, problems with grading, grading frameworks, and how to compute grades for nursing courses.

■ Purposes of Grades

In earlier chapters, there was extensive discussion about formative and summative evaluation. Through formative evaluation, the teacher provides feedback to the learner on a continuous basis. In contrast, summative evaluation is conducted periodically to indicate the student's achievement at the end of the course or at a point during the course. Summative evaluation provides the basis for arriving at grades in the course. *Grading*, or *marking*, is defined as the use of symbols, for instance, the letters A to F, for reporting student achievement. Grading is used for summative purposes, indicating through the use of symbols how well the student performed in individual assignments, clinical practice, laboratories (skills, simulation, others), and the course as a whole.

To reflect valid judgments about student achievement, grades need to be based on careful evaluation practices, reliable test results, and multiple assessment methods. No grade should be determined by one method or one assignment completed by the students; grades reflect instead a combination of various tests and other assessment methods. Along similar lines, students may complete assignments that are not included in their grade, particularly if the emphasis is on formative evaluation. Not all of the students' activities in a course need to be graded. Grades serve three broad purposes: (a) instructional, (b) administrative, and (c) guidance and counseling.

Instructional Purposes

Grades for instructional purposes indicate the achievement of students in the course. They provide a measure of *what* students have learned and their competencies at the end of the course or at a certain point within it. A “pass” grade in the clinical practicum and a grade of “B” in the nursing course are examples of using grades for instructional purposes.

Administrative Purposes

The second purpose that grades serve is administrative. Grades are used for:

- Admission of students to entry-level and higher degree nursing programs
- Progression of students in a nursing program
- Decisions about probation and whether students can continue in the program
- Decisions about reentry into a nursing program
- Determining students' eligibility for graduation
- Awarding scholarships and fellowships
- Awarding honors and determining acceptance into honor societies such as Sigma Theta Tau International
- Program evaluation studies
- Reporting competency to employers

Guidance and Counseling

The third use of grades is for guidance and counseling. Grades can be used to make decisions about courses to select, including more advanced courses to take or remedial courses that might be helpful. Grades also suggest academic resources that students might benefit from such as reading, study, and test-taking workshops and support. In some situations, grades assist students in making career choices, including a change in the direction of their careers.

■ Criticisms of Grades

Although grades serve varied purposes, there are some criticisms of them:

1. Grades are meaningless because of the diversity across nursing education programs, course faculty, clinical teachers, and preceptors.
 - Response: A consistent grading system is needed across sections of nursing courses and for grading clinical practice. It is important that

full- and part-time faculty members, clinical nurse educators, preceptors, and others involved in the course be oriented as to how to assess and grade each of the assignments. Clinical teachers and preceptors should understand the clinical evaluation process and methods, how to use the clinical evaluation tool, and grading practices in the course.

2. A single symbol, such as an A or a pass, does not adequately represent the complex details associated with achievement in nursing courses.
 - Response: Grades are not intended to fulfill this need. They do not reflect every aspect of the student's learning in a course or every accomplishment. Instead, grades are a summarization of achievements over a period of time.
3. Grades are not important.
 - Response: Although a grade is only a symbol of achievement, grades are important. Grades and overall grade point average (GPA) may predict later achievement such as performance on NCLEX® (National Council Licensure Examination) and certification examinations. Although some may argue that the most valuable outcomes of learning are intangible, grades, nevertheless, are important.
4. Self-evaluations are more important than grades.
 - Response: Developing the ability to evaluate one's own learning outcomes and competencies is essential for continued professional development. Both grades and self-evaluations are needed.
5. Grades are unnecessary.
 - Response: In most educational settings, grades cannot be eliminated because they serve the purposes identified earlier in the chapter. A certain level of performance is essential for progression in a nursing program and for later educational decisions; grades provide a way of determining whether students are achieving the outcome and competencies to progress through the program.
6. Grades are ineffective motivators.
 - Response: For many students, grades are effective motivators. In a study by Poorman and Mastorovich (2019), undergraduate, master's and doctoral nursing students all described the importance of grades to them. Students not only strived for an A overall, but they described the importance of a getting an A in every course they took. Grades to these nursing students were not only important but were motivators to work hard in their courses.
7. Low grades discourage students.
 - Response: Although low grades may be discouraging and stressful for students, they are essential for determining progression in a nursing

program. Nursing education programs are accountable to the profession and the public for preparing graduates with knowledge and competencies for safe practice. Not all entering students have the ability to acquire this knowledge and these skills. Low grades are important for counseling students and suggesting remedial instruction; failing grades indicate that students have not met the criteria for continuing in the nursing program.

8. Grades are inflated and thus do not reflect true achievement.

- Response: There has been considerable grade inflation over the past few decades. Students are paying more for their education, and they want a reward of high grades for their “purchase.” Other factors contributing to grade inflation are faculty hesitancy in assigning lower grades and not recognizing the problem with grade inflation; a close relationship of the teacher with students especially in clinical teaching, resulting in educators being more lenient in their grades; a lack of precision in tools used for evaluation; and grading systems in nursing education programs in which a C grade may be the lowest passing grade (Seldomridge & Walsh, 2018). In developing a grading system, it is important for nurse educators to be clear about the standards for each grade level in that system and to communicate these to students. Faculty also should periodically review the grades in nursing courses to assess whether they are inflated, keeping in mind that nursing students are carefully selected for admission into the program and need to achieve certain grades in courses to progress. For this reason, grades in nursing courses tend to be higher than general education courses in which students are more heterogeneous.

■ Types of Grading Systems

There are different types of grading systems or methods of reporting grades. Most nursing education programs use a letter system for grading (A, B, C, D, E or A, B, C, D, F), which may be combined with “+” and “−.” The integers 5, 4, 3, 2, and 1 (or 9–1) also may be used. These two systems of grading are convenient to use, yield grades that are able to be averaged within a course and across courses, and present the grade concisely.

Grades also may be indicated by percentages (100, 99, 98, ...). Most programs use percentages as a basis for assigning letter grades—90% to 100% represents an A, 80% to 89% represents a B, and so forth. In some nursing programs, the percentages for each letter grade are higher, for example, 93% to 100% for an A, 85% to 92% for a B, 76% to 84% for a C, 67% to 75% for a D, and 67% and below for an E or F. It is not

uncommon in nursing education programs to specify that students need to achieve at least a C in each nursing course at the undergraduate level and a B or better at the graduate level. Requirements such as these are indicated in the school policies and course syllabi.

Another type of grading system is two-dimensional: pass–fail, satisfactory–unsatisfactory, credit–no credit, and met–not met. For determining clinical grades, some programs add a third honors category, creating three levels: honors–pass–fail. One advantage of a two-dimensional grading system is that the grade is not calculated in the GPA. This allows students to take new courses and explore different areas of learning without concern about the grades in these courses affecting their overall GPA. This also may be viewed as a disadvantage, however, in that clinical performance in a nursing course graded on a pass–fail basis is not calculated as part of the overall course grade. A pass indicates that students met the outcomes or demonstrated satisfactory performance of the clinical competencies. Different systems for grading clinical practice are discussed later in the chapter.

Grade Point Average

One other dimension of a grading system involves converting the letter grade to a grade point system for calculating the GPA or quality point average (QPA). Grades in a 4-point system are typically:

A = 4 points per credit (or unit)

B = 3 points per credit

C = 2 points per credit

D = 1 point per credit

F = 0 points per credit

If a student took two 3-credit courses and one 6-credit nursing course and received an A in one of the 3-credit courses, a C in the other, and a B in the 6-credit course, the GPA would be

$A = 4 \text{ points/credit} = 4 \text{ points} \times 3 \text{ credits} = 12 \text{ points}$

$C = 2 \text{ points/credit} = 2 \text{ points} \times 3 \text{ credits} = 6 \text{ points}$

$B = 3 \text{ points/credit} = 3 \text{ points} \times 6 \text{ credits} = \underline{18 \text{ points}}$

$36 \div 12 (\text{credits}) = 3.0$

The letter system for grading also may include plus and minus grades. This is shown in Table 17.1. Although grade inflation may not decrease when plus and minus are used, these added categories allow for more differentiation for grading.

TABLE 17.1 Plus and Minus System

LETTER GRADE	GRADE POINTS
A	4.00
A–	3.67
B+	3.33
B	3.00
B–	2.67
C+	2.33
C	2.00
C–	1.67
D+	1.33
D	1.00
D–	0.67
F	0.00

■ Assigning Letter Grades

Because most nursing education programs use the letter system for grading nursing courses, this framework will be used for discussing how to assign grades. These principles, however, are applicable to the other grading systems as well. There are two major considerations in assigning letter grades: deciding what to include in the grade and selecting a grading framework.

Deciding What to Include in the Grade

Grades in nursing courses should reflect the student's achievement and not be biased by the teacher's own values, beliefs, and attitudes. If the student did not attend class or appeared to be inattentive during lectures, this behavior should not be incorporated into the course grade unless criteria were established at the outset for class attendance and participation.

The student's grade is based on the tests and assessment methods developed for the course. Multiple assessment methods should be used to determine course grades. The weight given to each of these in the overall grade should reflect the emphasis of the objectives and the content measured by them. Tests and other assessment methods associated with important content, for which more time was probably spent in the instruction, should receive greater weight in the course grade. For example, a midterm examination in a community health nursing course should be given more

weight in the course grade than a short paper that students completed about community resources for a family under their care.

How much weight should be given in the course grade to each test and other types of assessment methods used in the course? The teacher begins by listing the tests, quizzes, papers, presentations, and other assessment methods in the course that should be included in the course grade. Then the teacher decides on the importance of each of these components in the overall grade for the course. Factors to consider when weighting the components of the course grade are as follows:

1. Components that assess more of the important learning outcomes and competencies should carry more weight in the course grade than those that assess only a few of the outcomes.
2. Components that assess content that was emphasized in the course and for which more time was spent in the instruction and learning activities should receive the most weight in the course grade.
3. Components that measure the application of concepts and knowledge to clinical practice and development of higher level skills should be weighted more heavily than those that focus on recall of content.
4. Components that are more difficult and time-consuming for students should receive more weight than those that are easy and require less time to complete.

Selecting a Grading Framework

To give meaning to the grades assigned, the teacher needs a grading framework or frame of reference. There are three grading frameworks used to assign meaning to grades:

1. Criterion-referenced, also referred to as grading with an absolute scale
2. Norm-referenced or grading with a relative scale
3. Self-referenced or grading based on the growth of the student (Brookhart & Nitko, 2019)

Table 17.2 illustrates these grading frameworks. Criterion- and norm-referenced evaluation methods were described in earlier chapters; these same concepts apply to grading frameworks.

■ Criterion-Referenced Grading

In criterion-referenced grading, grades are based on students' achievement of the outcomes of the course, the extent of content learned in the course, or how well they performed in clinical practice. Students who achieve more of the objectives, acquire

TABLE 17.2 Grading Frameworks

GRADE	CRITERION-REFERENCED	NORM-REFERENCED	SELF-REFERENCED
A	All outcomes met Significant knowledge and skills gained Able to perform all clinical competencies	Achievement or performance far exceeds average of group (e.g., other students in course, in clinical group)	Made significant progress Performed significantly better than expected
B	Met all essential outcomes and at least half of the others Important content areas learned and able to be applied to new situations Able to perform most clinical competencies	Above the average of the group	Made progress and gained knowledge and skills Performed better than expected
C	All essential outcomes met; learned essential content Able to perform essential clinical competencies	Average in comparison with the group	Made progress in most areas Met expected performance level
D	Only some essential outcomes met; limited understanding of content Unable to perform some essential clinical competencies	Below the average of the group	Made some gains Did not meet level of performance for which capable
F	Most outcomes not achieved; limited content learned Most clinical competencies not able to be performed	Failing achievement or performance in comparison with the group	Made no gains Performance significantly below capability

Note: Content of this table based on ideas in Brookhart and Nitko (2019).

more knowledge, and can perform more competencies or with greater proficiency receive higher grades. The meaning assigned to grades, then, is based on these absolute standards without regard to the achievement of other students. Using this frame of reference for grading means that it is possible for all students to achieve an A or a B in a course, if they meet the standards, or a D or F if they do not.

This framework is appropriate for most nursing courses because it focuses on outcomes and competencies to be achieved in the course. Criterion-referenced grading indicates how students are progressing toward meeting those outcomes (formative evaluation) and whether they have achieved them at the end of the course (summative evaluation). Norm-referenced grading, in contrast, is not appropriate for use in nursing courses that are based on standards or learning outcomes because it focuses on comparing students with one another, not on how they are progressing or on their achievement. For example, formative evaluation in a norm-referenced framework would indicate how each student ranks among the group rather than provide feedback on student progress in meeting the outcomes of the course and strategies for further learning.

Fixed-Percentage Method

There are several ways of assigning grades using a criterion-referenced system. One is called the *fixed-percentage method*. This method uses fixed ranges of percent-correct scores as the basis for assigning grades (Miller, Linn, & Gronlund, 2013). A common grading scale is 93% to 100% for an A, 85% to 92% for a B, 76% to 84% for a C, 67% to 75% for a D, and below 67% for an E or F. Each component of the course grade—written tests, quizzes, papers, case presentations, and other assignments—is given a percentage-correct score or percentage of the total points possible. For example, the student might have a score of 21 out of 25 on a quiz, or 84%. The component grades are then weighted, and the percentages are averaged to get the final grade, which is converted to a letter grade at the end of the course. With all grading systems, the students need to be informed as to how the grade will be assigned. If the fixed-percentage method is used, the students should know the scale for converting percentages to letter grades; this should be in the course syllabus with a clear explanation of how the course grade will be determined.

Computing a Composite (Single) Score for a Course

In using the fixed-percentage method, the first step, which is an important one, is to assign weights to each of the components of the grade. For example:

Paper on nursing interventions	10%
Papers critiquing issues in clinical practice	20%
Quizzes	10%
Midterm examination	20%
Electronic portfolio	20%
Final examination	<u>20%</u> 100%

TABLE 17.3 Fixed-Percentage Method for Grading Nursing Courses

COMPONENT OF COURSE GRADE				WEIGHT (%)		
Paper on nursing interventions				10		
Papers critiquing issues in clinical practice				20		
Quizzes				10		
Midterm examination				20		
Electronic portfolio				20		
Final examination				20		
Student	Intervention Paper (10%)	Issue Papers (20%)	Quizzes (10%)	Midterm (20%)	Portfolio (20%)	Final (20%)
Mary	85	94	98	92	94	91
Jane	76	78	63	79	70	79
Bob	82	86	89	81	80	83
Composite score for Mary:						
$[10(85) + 20(94) + 10(98) + 20(92) + 20(94) + 20(91)] \div 100^a = 92.5\%$						
Composite score for Jane:						
$[10(76) + 20(78) + 10(63) + 20(79) + 20(70) + 20(79)] \div 100 = 75.1\%$						
Composite score for Bob:						
$[10(82) + 20(86) + 10(89) + 20(81) + 20(80) + 20(83)] \div 100 = 83.1\%$						

Note: 100 = sum of weights.

In determining the composite score for the course, the student's percentage for each of the components of the grade is multiplied by the weight and summed; the sum is then divided by the sum of the weights. This procedure is shown in Table 17.3.

Generally, test and other component scores should not be converted to grades for the purpose of later computing a final average grade. Instead, the teacher should record actual test scores and then combine them into a composite score that can be converted to a final grade.

Total-Points Method

The second method of assigning grades in a criterion-referenced system is the *total-points method*. In this method, each component of the grade is assigned a specific number of points, for example, a paper may be worth 100 points and midterm examination 75 points. The number of points assigned reflects the weights given to each component within the course, that is, what each one is "worth." For example:

Paper on nursing interventions	75 points
Papers critiquing issues in clinical practice	100 points
Quizzes	50 points
Midterm examination	75 points
Electronic portfolio	100 points
Final examination	<u>100 points</u> 500 points

The points for each component are not converted to a letter grade; instead, the grades are assigned according to the number of total points accumulated at the end of the course. At that time, letter grades are assigned based on the points needed for each grade. For example:

GRADE	POINTS
A	465–500
B	425–464
C	380–424
D	335–379
F	0–334

One problem with this method is that the decision about the points to allot to each test and evaluation method in the course is made before the teacher has developed them (Brookhart & Nitko, 2019). For example, to end with 500 points for the course, the teacher may need 75 points for the midterm exam. However, in preparing that exam, the teacher finds that 73 items adequately cover the content and reflect the emphasis given to the content in the instruction. If this were known during the course planning, the teacher could assign 2 fewer points to the midterm exam and add 2 points to the final exam or one of the assignments, or merely alter the total number of points for the course grade. However, when the course is already underway, changes such as these cannot be made in the grading scheme, and the teacher needs to develop a 75-point midterm exam even if fewer items would have adequately sampled the content. The next time the course is offered, the teacher can modify the points allotted for the midterm exam in the course grade.

Computing a Composite Score for a Course

In this method, the composite score is the total number of points the student accumulates, and no further calculations are needed. It is important that the weights

of the components of the grade are reflected in the points given them in the total composite. For example, if the teacher wanted the electronic portfolio to count as 20% of the course grade, and the maximum number of points for the course was 500, then the portfolio would be worth a maximum of 100 points (= 20% of 500).

■ Norm-Referenced Grading

In a norm-referenced grading system using relative standards, grades are assigned by comparing a student's performance with that of others in the class. Students who perform better than their peers receive higher grades. When using a norm-referenced system, the teacher decides on the reference group against which to compare a student's performance. Should students be compared with others in the course? Should they be compared with students only in their section of the course? Or with students who completed the course the prior semester or previous year? One issue with norm-referenced grading is that high performance in a particular group may not be indicative of mastery of the content or what students have learned; it reflects instead a student's standing in that group.

Grading on the Curve

Two methods of assigning grades using a norm-referenced system are (a) "grading on the curve" and (b) using standard deviations. *Grading on the curve* refers to the score distribution curve. In this method, students' scores are rank-ordered from highest to lowest, and grades are assigned according to the rank order. For example, the teacher may decide on the following framework for grading a test:

Top 20% of students	A
Next 20%	B
Next 40%	C
Next 15%	D
Lowest 5%	F

With this method, there would always be failing grades on a test.

After the quotas are set, grades are assigned without considering actual achievement. For example, the top 20% of students will receive an A even if their scores are close to the next group that gets a B. The students assigned lower grades may in fact have acquired sufficient knowledge in the course but unfortunately had lower scores than the other students. In these two examples, the decisions on the percentages of As, Bs, Cs, and lower grades are made arbitrarily by the teacher. The teacher determines the proportion of grades at each level; this approach is not based on a normal curve.

Another way of grading on the curve is to use the normal or bell curve for determining the percentages of each grade. The assumption of this method is that the grades of the students in the course should reflect a normal distribution. For example:

Top 10% of students	A
Next 20%	B
Next 40%	C
Next 20%	D
Lowest 10%	F

For “grading on the curve” to work correctly, student scores need to be distributed based on the normal curve. However, the abilities of nursing students tend not to be heterogeneous, especially late in the nursing education program, and therefore their scores on tests and other evaluation products are not normally distributed. They are carefully selected for admission into the program, and they need to achieve certain grades in courses and earn minimum GPAs to progress in the program. With grading on the curve, even if most students achieved high grades on a test and mastered the content, some would still be assigned lower grades.

Standard Deviation Method

The second method is based on standard deviations. With this method, the teacher determines the cutoff points for each grade. The grades are based on how far they are from the mean of raw scores for the class. To use the standard deviation method, the teacher first prepares a frequency distribution of the final scores and then calculates the mean score. With this method, the teacher has a reference point (mean) and the average distance of scores from the mean (Miller et al., 2013). The grade boundaries are then based on the standard deviation. For example, the cutoff points for a C grade might range from one half the standard deviation below the mean to one half above the mean. To identify the A–B cutoff scores, the teacher adds one standard deviation to the upper cutoff number of the C range. Subtracting one standard deviation from the lower C cutoff provides the range for the D–F grades.

■ Self-Referenced Grading

Self-referenced grading is based on standards of growth and change in the student. With this method, grades are assigned by comparing the student’s performance with the teacher’s perceptions of the student’s capabilities or the student’s own progress over the course (Brookhart & Nitko, 2019). Did the student achieve at a higher level

than deemed capable regardless of the knowledge and competencies acquired? Did the student improve performance throughout the course?

Table 17.2 compares self-referencing with criterion- and norm-referenced grading. One major problem with this method is the unreliability of the teacher's perceptions of student capability and growth, and the student's own assessment of performance. A second issue occurs with students who enter the course or clinical practice with a high level of achievement and proficiency in many of the clinical competencies. These students may have the least amount of growth and change but nevertheless exit the course with the highest achievement and clinical competency. Ultimately, judgments about the quality of a nursing student's performance are more important than judgments about the degree of improvement. It is difficult to make valid predictions about future performance on licensure or certification exams, or in clinical practice, based on self-referenced grades. For these reasons, self-referenced grades are not widely used in nursing education programs.

■ Grading Clinical Practice

Arriving at grades for clinical practice is difficult because of the nature of clinical practice and the need for judgments about performance. Issues in evaluating clinical practice and rating performance were discussed in Chapter 13, Clinical Evaluation Process and Chapter 14, Clinical Evaluation Methods. Many teachers constantly revise their rating forms for clinical evaluation and seek new ways of grading clinical practice. Although these changes may create a fairer grading system, they will not eliminate the problems inherent in judging clinical performance.

The different types of grading systems described earlier may be used for grading clinical practice. In general, these include systems using letter grades, A to F; integers, 5 to 1; and percentages. Grading systems for clinical practice also may use pass–fail, satisfactory–unsatisfactory, and met–did not meet the clinical objectives. Some programs add a third category, honors, to acknowledge performance that exceeds the level required. Pass–fail is used most frequently in nursing programs (Oermann, Yarbrough, Ard, Saewert, & Charasika, 2009). With any of the grading systems, it is not always easy to summarize the multiple types of evaluation data collected on the student's performance in a symbol representing a grade. This is true even in a pass–fail system; it may be difficult to arrive at a judgment as to pass or fail based on the evaluation data and the circumstances associated with the student's clinical, simulated, and laboratory practice.

Regardless of the grading system for clinical practice, there are two criteria to be met: (a) the evaluation methods for collecting data about student performance should reflect the outcomes and clinical competencies for which a grade will be assigned, and (b) students must understand how their clinical practice will be evaluated and graded.

Decisions about assigning letter grades for clinical practice are the same as grading any course: identifying what to include in the clinical grade and selecting a grading framework. The first consideration relates to the evaluation methods used in the course to provide data for determining the clinical grade. Some of these evaluation methods are for summative evaluation, thereby providing a source of information for inclusion in the clinical grade. Other strategies, though, are used in clinical practice for feedback only and are not incorporated into the grade.

The second consideration is the grading framework. Will achievement in clinical practice be graded from A to F? 5 to 1? Pass–fail? Or variations of these? A related question is, How will the clinical grade be included in the course grade, if at all?

Pass–Fail

Categories for grading clinical practice, such as pass–fail, satisfactory–unsatisfactory, and met–not met, have some advantages over a system with multiple levels, although there are disadvantages as well. Pass–fail places greater emphasis on giving feedback to the learner because only two categories of performance need to be determined. With a pass–fail grading system, teachers may be more inclined to provide continual feedback to learners because ultimately they do not have to differentiate performance according to four or five levels of proficiency such as with a letter system. Performance that exceeds the requirements and expectations, however, is not reflected in the grade for clinical practice unless a third category is included: honors–pass–fail.

A pass–fail system requires only two types of judgment about clinical performance. Do the evaluation data indicate that the student has met the outcomes or has demonstrated satisfactory performance of the clinical competencies to indicate a pass? Or do the data suggest that the performance of those competencies is not at a satisfactory level? Arriving at a judgment as to pass or fail is often easier for the teacher than using the same evaluation information for deciding on multiple levels of performance. Use of a letter system for grading clinical practice, however, acknowledges the different levels of clinical proficiency students may have demonstrated in their clinical practice.

A disadvantage of pass–fail for grading clinical practice is the difficulty of including a clinical grade in the course grade. One strategy is to separate nursing courses into two components for grading, one for theory and another for clinical practice (designated as pass–fail), even though the course may be considered as a whole. Typically, guidelines for the course indicate that the students must pass the clinical component to pass the course. An alternative mechanism is to offer two separate courses with the clinical course graded on a pass–fail basis or by using a letter system.

Once the grading system is determined, there are various ways of using it to arrive at the clinical grade. In one method, the grade is assigned based on the outcomes or

competencies achieved by the student. To use this method, the teacher should consider designating some of the outcomes or competencies as critical for achievement. Table 17.2 provides guidelines for converting the clinical competencies into letter grades within a criterion-referenced system. For example, an A might be assigned if all of the competencies were achieved; a B might be assigned if all of the critical ones were achieved and at least half of the others were met.

For pass–fail grading, teachers can indicate that all of the outcomes or competencies must be met to pass the course, or they can designate critical behaviors required for passing the course. In both methods, the clinical evaluation methods provide the data for determining whether the student’s performance reflects achievement of the competencies. These evaluation methods may or may not be graded separately as part of the course grade.

Another way of arriving at the clinical grade is to base it on the evaluation methods. In this system, the clinical evaluation methods become the source of data for the grade. For example:

Paper on analysis of clinical practice issue	10%
Analysis of clinical cases	5%
Conference presentation	10%
Community resource paper	10%
Electronic portfolio	25%
Rating scale (of performance)	40%

In this illustration, the clinical grade is computed according to the evaluation methods. Observation of performance, and the rating on the clinical evaluation tool, comprise only a portion of the clinical grade. An advantage of this approach is that it incorporates into the grade the summative evaluation methods completed by the students.

If pass–fail is used for grading clinical practice, the grade might be computed as follows:

Paper on analysis of clinical practice issue	10%
Analysis of clinical cases	5%
Conference presentation	10%
Community resource paper	10%
Electronic portfolio	25%
OSCE	40%
Rating scale (of performance)	Pass required

OSCE, objective structured clinical examination.

This discussion of grading clinical practice has suggested a variety of mechanisms that are appropriate. The teacher must make it clear to the students and others how the evaluation and grading will be carried out.

Failing Clinical Practice

Teachers will be faced with determining when students have not met the outcomes of clinical practice, that is, have not demonstrated sufficient competence to pass the clinical course. There are principles that should be followed in evaluating and grading clinical practice, which are critical if a student fails a clinical course or has the potential to fail it.

Communicate Evaluation and Grading Methods in Writing

The evaluation methods used in a clinical course, the manner in which each will be graded if at all, and how the clinical grade will be assigned should be documented in writing and communicated to the students. The practices of the teacher in evaluating and grading clinical performance must reflect this written information. In courses with preceptors, it is critical that preceptors and others involved in teaching and assessing student performance understand the outcomes of the course, the evaluation methods, how to observe and rate performance, and the preceptor's responsibilities when students are not performing adequately. Preceptors are reluctant to assign failing grades to students whose competence is questionable (Anthony & Wickman, 2015). There is a need for faculty development, especially for new and part-time teachers. As part of this education teachers should explore their beliefs and values about grading clinical performance and their expectations of students in the clinical setting.

Identify Effect of Failing Clinical Practicum on Course Grade

If failing clinical practice, whether in a pass–fail or a letter system, means failing the nursing course, this should be stated clearly in the course syllabus and policies. By stating it in the syllabus, which all students receive, the students have it in writing before clinical learning activities begin. A sample policy statement for pass–fail clinical grading is:

The clinical component of NUR XXX is evaluated with a grade of pass or fail. A fail in the clinical component results in failure of the course even if the theory grade is 75% or higher.

In a letter-grade system, the policy should include the letter grade representing a failure in clinical practice, for example, less than a C. A sample policy statement for this system is:

Students must pass the clinical component of NUR XXX with the grade of “C” or higher. A grade lower than a “C” in the clinical component of the course results in failure of the course even if the theory grade is 75% or higher.

Ask Students to Sign Notes, Rating Forms, and Evaluation Summaries

Students should sign any written clinical evaluation documents—notes about the student's performance in clinical practice, rating forms (of clinical practice, clinical examinations, and performance in simulations), narrative comments about the student's performance, and summaries of conferences in which performance was discussed. Their signatures do not mean they agree with the ratings or comments, only that they have read them. Students should have an opportunity to write in their own comments. These materials are important because they document the student's performance and indicate that the teacher provided feedback and shared concerns about that performance. This is critical in situations in which students may be failing the clinical course because of performance problems.

Identify Performance Problems Early and Develop Learning Contracts

Students need continuous feedback on their clinical performance. Observations made by the teacher, the preceptor, and others, as well as evaluation data from other sources, should be shared with the student. Performance data should be discussed together. Students may have different perceptions of their performance and in some cases may provide new information that influences the teacher's judgment about clinical competencies.

When the teacher or preceptor identifies performance problems and clinical deficiencies that may affect passing the course, conferences should be held with the student to discuss these areas of concern and develop a plan for remediation. It is critical that these conferences focus on problems in performance combined with specific learning activities meant to address them. The conferences should not consist of the teacher telling the student everything that is wrong with his or her clinical performance; the student needs an opportunity to respond to the teacher's concerns and identify how to address them.

One of the goals of the conference is to develop a plan with learning activities for the student to correct deficiencies and develop competencies further. This plan serves as a learning contract, an agreement between the teacher and student. A learning contract specifies the clinical competencies to be developed by the student, learning activities planned collaboratively by the teacher and student to guide learning and improve performance, and expected outcomes with "due dates." Exhibit 17.1 is an example of a format that can be used to develop a learning contract for any level of learner. If the student is failing clinical practice, the contract should indicate that (a) completing the remedial learning activities does not guarantee that the student will pass the course, (b) one satisfactory performance of the competencies will not constitute a pass clinical grade (the improvement must be sustained), and (c) the student must demonstrate satisfactory performance of the competencies by the end of the course.

EXHIBIT 17.1**LEARNING CONTRACT TEMPLATE***Student Information: Name, Course, Contact Information**Teacher Information: Name, Course, Contact Information*

COURSE OUTCOMES OR CLINICAL COMPETEN- CIES TO BE ACHIEVED	REQUIRED LEARNING ACTIVITIES (WITH RESOURCES FOR LEARN- ING)	DUE DATE FOR COM- PLETION	EVALUATION EVIDENCE, RESPON- SIBILITY	DUE DATE FOR ACHIEVE- MENT OF OUTCOMES OR COMPE- TENCIES
1.				
2.				
3.				
4.				

Start Date: _____

Completion Date: _____

Student Signature: _____

Teacher Signature: _____

Preceptor Signature (if applicable): _____

Any discussions with students at risk of failing clinical practice should focus on the student's inability to achieve the outcomes of the clinical course and perform the specified competencies, not on the teacher's perceptions of the student's intelligence, overall ability, or perceived motivation or effort. In addition, opinions about the student's ability in general should not be discussed with others.

Conferences should be held in private, and a summary of the discussion should be prepared. The summary should include the date and time of the conference, who participated, areas of concern about clinical performance, and the learning plan with a time frame for completion (Oermann, Shellenbarger, & Gaberson, 2018). The summary should be signed by the teacher, the student, and any other participants. Faculty members should review related policies of the nursing education program because they might specify other requirements.

Identify Support Services

Students at risk for failing clinical practice may have other problems that are affecting their performance. Teachers should refer students to counseling and other support services and not attempt to provide these resources themselves. Attempting to

counsel the student and help the student cope with other problems may bias the teacher and influence judgment of the student's clinical performance.

Document Performance

As the clinical course progresses, the teacher should give feedback to the student about performance and continue to guide learning. It is important to document the observations made, other types of evaluation data collected, and the learning activities completed by the student. The documentation should be shared routinely with students, discussions about performance should be summarized, and students should sign these summaries to confirm that they read them.

The teacher cannot observe and document the performance *only* of the student at risk for failing the course. There should be a minimum number of observations and documentation of other students in the clinical group, or the student failing the course might believe that he or she was treated differently than others in the group. One strategy is to plan an approximate number of observations of performance to be made for each student in the clinical group to avoid focusing only on the student with performance problems. However, teachers may observe students who are believed to be at risk for failure more closely, and document their observations and conferences with those students more thoroughly and frequently than is necessary for the majority of students. When observations result in feedback to students that can be used to improve performance, at-risk students usually do not object to this extra attention.

Follow Policy on Unsafe Clinical Performance

There should be a policy in the nursing program about actions to be taken if a student's work in clinical practice is unsafe. Students who are not meeting the outcomes of the course or have problems performing some of the competencies can continue in the clinical course as long as they demonstrate safe care. This is because the outcomes and clinical competencies are identified for achievement at the *end* of the course, not during it.

If the student demonstrates performance that is potentially unsafe, however, the teacher can remove the student from the clinical setting when following the policy and procedures of the nursing education program. Specific learning activities outside of the clinical setting need to be offered to help students develop the knowledge and skills they lack; simulation and practice in the skills laboratory are valuable in these situations. A learning plan should be prepared and implemented as described earlier.

Follow Policy for Failure of a Clinical Course

In all instances, the teacher must follow the policies of the nursing program. If the student fails the clinical course, the student must be notified of the failure and its consequences as indicated in these policies. In some nursing education programs,

students are allowed to repeat only one clinical course, and there may be other requirements to be met. If the student will be dismissed from the program because of the failure, the student must be informed of this in writing. Generally, there is a specific time frame outlined for each step in the process, which must be adhered to by the faculty, administrators, and students. It is critical that all teachers know the policies and procedures to be implemented when students have performance problems or are at risk for failing the clinical course. These policies and procedures must be followed for all students.

■ Grading Software

A number of the procedures used to determine grades are time-consuming to use, particularly if the class of students is large. Although a calculator may be used, student grades can be calculated easily with a spreadsheet application such as Microsoft Excel or in an online learning management system. With a spreadsheet application, teachers can enter individual scores, include the weights of each component of the grade, and compute final grades. Many statistical functions can be performed with a spreadsheet application.

Learning management systems provide grade books for teachers to manage all aspects of student grades. The grades can be weighted and a final grade calculated. One advantage to using a learning management system grade book is that students usually have online access to their own scores and grades as soon as the teacher has entered them.

There are also a number of grading software programs on the market that include a premade spreadsheet for grading purposes; these have different grading frameworks that may be used to calculate the grade and enable the teacher to carry out the tasks needed for grading. Not all grading software programs are of high quality, however, and should be reviewed prior to purchase.

■ Summary

Grading is the use of symbols, such as the letters A to F, to report student achievement. Grading is used for summative purposes, indicating how well the student met the outcomes of the course and performed in clinical practice. Grades need to be based on careful evaluation practices, valid and reliable test results, and multiple assessment methods. No grade should be determined on the basis of one method or one assignment completed by the students; grades reflect instead a combination of various tests and other assessment methods.

There are different types of grading systems or methods of reporting grades: the use of letters A to E or A to F, which may be combined with “+” and “-”; integers 5, 4, 3, 2, and 1 (or 9–1); percentages; and categories such as pass–fail and

satisfactory–unsatisfactory. Advantages and disadvantages of pass–fail for grading clinical practice were discussed in the chapter.

Two major considerations in assigning letter grades are deciding what to include in the grade and selecting a grading framework. The weight given to each test and the evaluation method in the grade is specified by the teacher according to the emphasis of the course outcomes and the content measured by them. To give meaning to the grades assigned, the teacher needs a grading framework: criterion referenced, also referred to as *grading with absolute standards*; norm referenced, or *grading with relative standards*; or self-referenced, *grading based on the growth of the student*.

One final concept described in the chapter was grading clinical practice and guidelines for working with students who are at risk for failing a clinical course. These guidelines give direction to teachers in establishing sound grading practices and following them when working with students in clinical practice.

■ References

- Anthony, M. L., & Wickman, M. (2015). Precepting challenges: The unsafe student. *Nurse Educator*, 40, 113–114. doi:10.1097/NNE.0000000000000118
- Brookhart, S. M., & Nitko, A. J. (2019). *Educational assessment of students* (8th ed.). New York, NY: Pearson Education.
- Miller, M. D., Linn, R. L., & Gronlund, N. E. (2013). *Measurement and assessment in teaching* (11th ed.). Upper Saddle River, NJ: Pearson Education.
- Oermann, M. H., Shellenbarger, T., & Gaberson, K. B. (2018). *Clinical teaching strategies in nursing* (5th ed.). New York, NY: Springer Publishing Company.
- Oermann, M. H., Yarbrough, S. S., Ard, N., Saewert, K. J., & Charasika, M. (2009). Clinical evaluation and grading practices in schools of nursing: Findings of the evaluation of learning advisory council survey. *Nursing Education Perspectives*, 30, 352–357.
- Poorman, S. G., & Mastorovich, M. L. (2019). The meaning of grades: Stories of undergraduate, master's, and doctoral nursing students. *Nurse Educator*. doi:10.1097/nne.0000000000000627. [e-pub ahead of print]
- Seldomridge, L. A., & Walsh, C. M. (2018). Waging the war on clinical grade inflation: The ongoing quest. *Nurse Educator*, 43, 178–182. doi:10.1097/nne.0000000000000473

PROGRAM EVALUATION AND ACCREDITATION

Program evaluation is the process of judging the worth or value of an educational program. One purpose of program evaluation is to provide data on which to base decisions about the educational program. Another purpose is to provide evidence of educational effectiveness in response to internal and external demands for accountability. With the demand for high-quality programs, development of newer models for the delivery of higher education such as online programs, and public calls for accountability, there has been a greater emphasis on systematic and ongoing program evaluation. This chapter presents an overview of accreditation of nursing education programs, standards for evaluating distance education programs, program evaluation models, and the evaluation of selected program components (curriculum, outcomes, and teaching). The chapter also describes how to develop a program evaluation plan for a nursing education program.

■ Accreditation of Nursing Education Programs

Accreditation is a means of ensuring that institutions of higher education and nursing education programs meet standards and of indicating to the public that they offer a quality education for students. In this way, accreditation protects the public. Accreditation involves internally reviewing and assessing one's own programs on a continuous basis to identify areas requiring revision and participating in an external review, including a site visit, to verify the standards are met.

Nursing accreditation is not the same as regulation of nursing education programs. The state boards of nursing are regulatory bodies: They set standards for nursing practice and education in a state. Nursing education programs cannot operate without approval from the state board (Halstead, 2017). Boards of nursing, which are governmental agencies, have this authority because they have responsibility in the state to protect the health and welfare of their citizens by overseeing and ensuring safe nursing practice (National Council of State Boards of Nursing, 2019). Each state board has its own set of standards that nursing education programs within the state need to meet for initial and continuing approval. State boards have the authority

to close a program or require improvements to continue to operate, for example, if NCLEX® (National Council Licensure Examination) scores or retention rates are below the state's benchmarks for a designated period of time.

Accreditation is a voluntary process done by nongovernmental agencies. Although accreditation is voluntary, in many ways, it is difficult for nonaccredited nursing education programs to attract students because financial aid for students from both the federal and state governments is contingent on attending an accredited nursing education program.

Types of Accreditation

There are two types of accreditation: institutional (at the university or college level) and programmatic (at the program or discipline level). Institutional accreditation is done by regional accrediting bodies such as the Southern Association of Colleges and Schools Commission on Colleges, and the Middle States Commission on Higher Education. Each region of the United States has its own regional accrediting body. At the institutional level, there also are accrediting bodies for faith-based colleges and career-related schools and programs.

Programmatic or specialized accreditation is performed for a field of study, for example, nursing, business, engineering, and medicine, among others. Programmatic accreditation is typically found in the professions, meeting their responsibility to ensure quality of programs to protect the public. There are three accrediting bodies for nursing education programs in the United States: Accrediting Commission for Education in Nursing (ACEN), Commission on Collegiate Nursing Education (CCNE), and Commission for Nursing Education Accreditation (CNEA). Two of these agencies (ACEN and CNEA) accredit all levels of nursing education programs: practical, diploma, associate, baccalaureate, masters, and DNP (ACEN, 2019b; CNEA, 2019). CCNE accredits baccalaureate and master's degree nursing programs, practice-focused nursing doctoral programs that award the DNP degree, postgraduate APRN certificate programs, and entry-to-practice residency programs (CCNE, 2019). The Canadian Association of Schools of Nursing serves as the national accrediting body for RN education in Canada. In specialty areas of nursing practice, such as nurse anesthesia, there are specialty accrediting bodies (e.g., Council on Accreditation of Nurse Anesthesia).

PhD in nursing programs are not accredited by any of these agencies because PhD (and other research-focused doctoral) programs usually are offered through the graduate school of the university. They are evaluated using processes determined by the university, which usually include an internal review and self-study, followed by an external review and site visit, similar to nursing accreditation processes.

Program evaluation based on an accreditation model is designed to assess whether the program meets external standards of quality. Although there are differences in the

names of the standards, the areas evaluated and descriptions of quality of nursing education programs in those areas are similar. Table 18.1 lists the areas addressed by the standards for the three accrediting bodies for nursing education. The focus is on quality improvement—collecting information about the program and its components on a continuous basis to identify what is working well and decide what changes may be needed.

TABLE 18.1 Areas of Accreditation Standards for Nursing Education Programs

AREAS	ACEN	CCNE	CNEA
Mission and governance	1. Mission and administrative capacity	1. Program quality: Mission and governance	2. Mission, governance, resources
Faculty	2. Faculty and staff	2. Program quality: Institutional commitment and resources (includes faculty)	3. Faculty
Students	3. Students	In multiple standards	4. Students
Curriculum	4. Curriculum	3. Program quality: Curriculum and teaching–learning practices	5. Curriculum and evaluative processes
Resources	5. Resources	2. Program quality: Institutional commitment and resources	2. Mission, governance, resources
Outcomes	6. Outcomes	4. Program effectiveness: Assessment and achievement of program outcome	1. Program outcomes
Available at	http://www.acenursing.net/manuals/SC2017.pdf	https://www.aacnnursing.org/Portals/42/CCNE/PDF/Standards-Final-2018.pdf	http://www.nln.org/docs/default-source/accreditation-services/cnea-standards-final-february-201613f2bf5c78366c709642ff00005f0421.pdf?sfvrsn=12

Note: Numbers indicate the number of the standard.

ACEN, Accreditation Commission for Education in Nursing; CCNE, Commission on Collegiate Nursing Education; CNEA, Commission for Nursing Education Accreditation.

Evaluation of Distance Education Programs Using an Accreditation Model

Although no specific accreditation standards exist for online nursing programs, ACEN (2019a), CCNE (2018), and CNEA (2016) incorporate distance education in their standards. For ACEN accreditation, distance education nursing programs must demonstrate compliance with 10 critical elements in addition to the appropriate accreditation standards for the level of program:

- Congruence with the institutional mission
- Instructional design and course delivery methods
- Competence and preparation of faculty
- Quality and accessibility of support services for students
- Quality and accessibility of support services for faculty members
- Current, relevant, and accessible learning resources
- Current, appropriate offerings relative to the delivery method
- Provision of opportunities for regular faculty–student and student–student interactions
- Ongoing evaluation of student learning
- Processes established for verifying student identity in courses (ACEN, 2019a)

The CCNE accreditation process incorporates specific elements related to effective distance education in three of the four accreditation standards, as the following examples illustrate:

- *Standard I, Program Quality: Mission and Governance:* Roles of the faculty and students in program governance, including participants in distance education, are clearly defined and promote participation.
- *Standard II, Program Quality: Institutional Commitment and Resources:* Academic support services, such as library, technology, distance education support, research support, admission, and advising services, support achievement of program outcomes, and there is a defined process for review.
- *Standard III, Program Quality: Curriculum and Teaching–Learning Practices:* Teaching–learning practices in all environments, including distance education, support student achievement of expected learning outcomes; clinical

practice experiences are provided for all students, including those in distance education offerings (CCNE, 2018).

The CNEA Standards of Accreditation are applicable to all types of nursing programs, including distance education programs (CNEA, 2016). In some of the standards, specific statements are included to guide faculty and administrators in how the standards apply to online programs. For example, one statement specifies that governance structures in the school should facilitate including distance education students.

In addition to applying the accreditation standards, the Council of Regional Accrediting Commissions (C-RAC) developed guidelines that could be used for evaluating online nursing programs: *Interregional Guidelines for the Evaluation of Distance Education* (C-RAC, 2011). There are nine guidelines, each with specific areas to address. These are summarized in Exhibit 18.1.

EXHIBIT 18.1

INTERREGIONAL GUIDELINES FOR EVALUATION OF DISTANCE EDUCATION

1. Online programs should be appropriate for the mission and goals of the institution.
2. There are plans for developing, sustaining, and, if appropriate, expanding online offerings, which are part of the regular evaluation processes.
3. Online learning is incorporated into the governance and academic oversight of the institution.
4. Curricula are “coherent, cohesive, and comparable in academic rigor” to traditional programs offered by the institution.
5. There is an evaluation of online learning offerings with its results used for improvement.
6. The faculty is qualified and supported.
7. The institution provides effective student and academic support services.
8. There are sufficient resources to support and expand (if indicated) online course offerings.
9. The institution assures the integrity of online offerings (C-RAC, 2011).

There also are standards from the Distance Education Accrediting Commission (DEAC). Each of these standards, listed in Exhibit 18.2, includes a description of quality practices and expectations. These areas are applicable to assessing online programs in nursing.

EXHIBIT 18.2**DEAC ACCREDITATION STANDARDS**

Standards for evaluating distance education programs relate to:

- Institutional mission, effectiveness, and strategic planning
- Program outcomes, curriculum, and supplemental materials
- Educational and student support services
- Assessment plan to track student achievement and satisfaction
- Academic leadership and faculty qualifications
- Advertising and promotion of the institution and programs, and recruitment personnel
- Admission criteria and practices, and enrollment agreements
- Financial disclosures, cancellations, and refunds
- Institutional governance
- Financial responsibilities
- Facilities and supplies (including record keeping; DEAC, 2019)

DEAC, Distance Education Accrediting Commission.

■ Program Evaluation Models

Saewert (2017) described eight models for program evaluation: objective-based, goal-free, expert-oriented, naturalistic, participative oriented, improvement-focused, success case, and theory-driven. Each of these models has positive and negative considerations for its use. Nurse educators should seek a model that will help organize the program evaluation and produce the most useful information for various stakeholders. Some nursing education programs use an eclectic approach in which they design their own model by selecting features from more than one (Saewert, 2017).

Another type of model is decision oriented. With these models, the goal of the evaluation is to provide information to decision makers for program improvement purposes. However, the existence of assessment data is no guarantee that the program will use the data to improve (Stufflebeam, 2013). Decision models focus more on using assessment data as a tool to improve programs than on accountability. Decision-oriented models usually focus on internal standards of quality, value, and efficacy. An example of a decision-oriented model is Stufflebeam's Context, Input, Process, Product (CIPP) Model (Stufflebeam, 2013).

Other models are systems oriented. These examine inputs into the program such as characteristics of students, teachers, administrators, and other participants in the program, as well as program resources. These models also assess the operations and processes of the program as well as the context or environment within which

the program is implemented. Finally, systems models examine the outcomes of the program: Are the intended outcomes being achieved? Are students, graduates, their employers, faculty, staff, and others satisfied with the program and how it is implemented? Is the program of high quality and cost-effective?

Regardless of the specific model used, the process of program evaluation assists various audiences or stakeholders of an educational program in judging and improving its worth or value. Audiences or stakeholders are those individuals and groups who are affected directly or indirectly by the decisions made. Key stakeholders of nursing education programs include students, faculty and staff members, partners (healthcare and community agencies), employers of graduates, and consumers. The purpose of the program evaluation determines which audiences should be involved in generating questions or concerns to be answered or addressed. When the focus is formative, that is, to improve the program during its implementation, the primary audiences are students, teachers, and administrators. Summative evaluation leads to decisions about whether a program should be continued, revised, funded, or terminated. Audiences for summative evaluation include program participants, graduates, their employers, prospective students, healthcare and community agencies, consumers, legislative bodies, funding agencies, and others who might be affected by changes in the program.

When planning program evaluation, an important decision is whether to use external or internal evaluators. External evaluators are thought to provide objectivity, but they may not know the program or its context well enough to be effective. External evaluators also add expense to the program assessment. In contrast, an internal evaluator has a better understanding of the operations and environment of the program and can provide continuous feedback to the individuals and groups responsible for the evaluation. However, an internal evaluator may be biased, reducing the credibility of the evaluation.

■ Curriculum Evaluation

Curriculum evaluation is not the same as program evaluation. When evaluating the curriculum, the focus is on elements central to the course of studies taken by students. As such, curriculum evaluation is narrower than program evaluation, which includes all the elements needed to offer a quality nursing program such as institutional support, qualified faculty, and adequate resources, among other areas. In evaluating the curriculum, the main areas to assess are:

- *Philosophy*: Is the philosophy current, and are the beliefs and values in the philosophy guiding teaching and decisions in the school?
- *Conceptual framework*: Is there a framework for the curriculum, and is it reflected in the program and student learning outcomes, courses, learning activities, and other aspects of the curriculum?

- *Program outcomes*: Do the program outcomes indicate important qualities and characteristics of the graduates on completion of the program, and are they student centered? Are they appropriate to the level of the nursing education program and relevant to the healthcare context? If the program has level objectives, are these congruent with the program outcomes, do they reflect student achievement at designated points in time (levels), and are they different from course outcomes?
- *Curriculum design*: How well do the curriculum components fit together? Do nursing courses build on foundational and prerequisite courses? Are the courses logically sequenced, and do they build on one another?
- *Courses*: Are course outcomes or objectives congruent with program and level outcomes? Are students achieving the outcomes of the course and developing essential knowledge, values, and skills for practice and progression through the curriculum? What is the learning environment in the course? Is the course content organized logically, and are exemplars used in the course relevant?
- *Teaching–learning strategies*: Do teaching strategies promote students' learning and achievement of course outcomes? Do they promote active learning? Do they respect student diversity? Are students satisfied with the teaching methods, learning activities, assignments, and quality of teaching in the course?
- *Assessment methods*: Are the assessment methods appropriate for the outcomes to be evaluated and level of students, and are they reasonable? Do they provide for demonstration of relevant types of learning? Do they accommodate students' diverse learning styles? Do they provide for feedback to improve students' learning and performance (formative) and periodic summative evaluation?
- *Resources and partnerships*: Are the resources (physical space, classrooms, technology, simulation, laboratories, clinical sites, etc.) sufficient to implement the curriculum? Are there adequate numbers of qualified faculty members to offer the curriculum and teach the courses? Are there well-prepared clinical instructors, preceptors, and other educators for clinical teaching? Are the number of staff members and their roles and functions appropriate for the curriculum? Are library holdings appropriate and sufficient, and are they available to students in online courses? Are there adequate and readily available support services for students including those in online courses? (Iwasiw, Andruszyn, & Goldenberg, 2020; Valiga, 2017).

Although curriculum evaluation is important, an educational program involves more than a curriculum. The success of students in meeting the outcomes of courses

and the curriculum as a whole may depend as much on the quality of the students admitted to the program or the characteristics of its faculty as it does on the sequence of courses or the instructional strategies used. Similarly, there may be abundant evidence that graduates meet the outcomes of the curriculum, but graduates may not be satisfied with the program or may be unable to pass licensure or certification examinations. As faculty review assessment data, they need to consider all elements of the program before making decisions about program changes.

Evaluation of Courses

Course evaluation is similar to curriculum evaluation because courses are elements of the curriculum. All aspects of course design and implementation should be included in the evaluation. Course evaluation may determine the extent to which:

- Course outcomes are appropriate and relate to the program outcomes.
- Courses reflect the background of students and whether they have prerequisite knowledge and skills.
- Students achieve learning outcomes in the course.
- Content is up-to-date and evidence based.
- The course is logically organized.
- Teaching methods and learning activities are relevant for the course outcomes (or objectives).
- Teaching methods and learning activities facilitate learning and engage students actively in the course.
- Assessment methods are related to the course outcomes.
- Assignments are appropriate and promote learning.
- Grading criteria are clear and adhered to.
- Students are satisfied with the course.
- Faculty are satisfied with the course.

Students typically are asked to evaluate their courses at the end of an academic term, but course evaluations need to extend beyond student ratings. Faculty members have the necessary in-depth knowledge of the curriculum to assess the value of each course in relation to intended curriculum outcomes.

Evaluation of Teaching–Learning Activities

When evaluating the teaching–learning activities used in a course, the faculty should give primary considerations as to whether they facilitated student learning and the

extent to which they engaged students actively. Teaching–learning strategies should be evidence based and reflect best practices (Oermann & Conklin, 2018; Valiga, 2017). Although student satisfaction with teaching–learning activities is important, it is insufficient. Faculty members should evaluate the effectiveness of the teaching methods and each assignment in a course to ensure it is the best approach for the outcomes to be achieved. Teaching methods should promote active involvement of students and interactions with others, one-to-one and in small groups. Students should be satisfied with the teaching methods and learning activities in a course, but students are not experts in education, and faculty members need to choose the most effective strategies for promoting student learning. Some of those strategies, such as problem-based learning, require more time for students and active involvement, which may affect students’ satisfaction and evaluation of the course. Student satisfaction, although important, is not the only factor to consider.

■ Teaching

Another area of evaluation involves appraising the effectiveness of the teacher. This addresses the quality of teaching in the classroom, online, in simulation, in skills laboratories, and in clinical settings. Evaluation of teaching also includes other dimensions of the teacher’s role, depending on the goals and mission of the nursing education program. These other roles may involve scholarship and research; service to the nursing program, college or university, community, and nursing profession; and clinical practice. It is beyond the scope of this book to examine these multiple dimensions, but a brief discussion is provided about assessing the quality of teaching in the classroom and the clinical setting.

The National League for Nursing (NLN) *Nurse Educator Core Competency* describes the knowledge, skills, and attitudes required for effectiveness in the role of nurse educator as identified in research-based literature (NLN, 2019). In the NLN’s *The Scope of Practice for Academic Nurse Educators* (2012), each competency is followed by tasks that further describe the competency. These competencies address the ability of the nurse educator to guide students’ learning, professional development, and socialization into the nursing role, at whatever level they are preparing for; use assessment and evaluation methods; develop, evaluate, and revise nursing programs and courses; engage in scholarship in the educator role; be a leader and change agent; and contribute to the school of nursing or other educational setting.

The research in nursing education suggests five qualities of effective teaching in nursing: (a) knowledge, (b) clinical competence, (c) teaching skill, (d) interpersonal relationships with learners, and (e) personal characteristics. These findings are consistent with studies about teacher effectiveness in other fields (Oermann, Shellenbarger, & Gaberson, 2018).

Knowledge

An effective teacher is an expert in the content area and is knowledgeable about the area in which teaching. Nurse educators need to keep current with nursing practice, new developments in their areas of expertise, and research. However, knowledge alone is not sufficient; the teacher must be able to communicate that knowledge to students, assist students in applying knowledge to patient care, and help them learn.

Clinical Competence

If teaching in the clinical setting, the teacher has to be competent in clinical practice (Oermann et al., 2018). The clinical competence of the teacher is one of the most important characteristics of effective clinical teaching in nursing (Lovric, Prlic, Zec, Puseljic, & Zvanut, 2015). The best teachers are expert clinicians who know how to care for patients, can make sound clinical judgments, have expert clinical skills, and can guide students in developing those skills.

In some nursing education programs, especially graduate programs for advanced practice nursing, faculty members are required to maintain certification in clinical practice. In programs without such a requirement, the clinical teacher should maintain clinical competence by regular clinical practice, attending continuing nursing education programs, reading current clinical literature, or other means.

Teaching

Although necessary, knowing what to teach is not sufficient. The teacher also needs to know how to teach. Competencies in teaching involve the ability to:

- Identify students' learning needs.
- Plan instruction.
- Present content effectively.
- Explain concepts clearly with exemplars that help students apply those concepts in patient care.
- Demonstrate procedures effectively.
- Be skilled in teaching online and promoting interaction in an online environment (if relevant).
- Use sound assessment practices.

The research suggests that the teacher's skills in evaluation are particularly important to teaching effectiveness. Evaluating learners fairly, having clear expectations and communicating those to students, correcting mistakes without embarrassing students, and giving immediate feedback are important teacher behaviors (Oermann et al., 2018).

Interpersonal Relations With Learners

Another important characteristic is the ability of the teacher to establish positive relationships with students as a group in the classroom, online environment, and clinical setting, and with students on an individual basis. Effective teaching qualities in this area include showing respect for and confidence in students, being honest and direct, supporting students, and being approachable (Oermann et al., 2018). Effective teachers treat students fairly and create an environment of mutual respect between educator and student.

Personal Characteristics of Teacher

Effective teaching also depends on the personal characteristics of the teacher. Characteristics in this area include enthusiasm, patience, having a sense of humor, friendliness, integrity, perseverance, courage, and willingness to admit mistakes (Oermann et al., 2018). Personal characteristics that convey caring about students support caring as a core value in nursing. Caring about students, recognizing they are learners not yet nurses, and supporting them as they learn and develop their skills are particularly important in clinical teaching, which is stressful for students.

■ How to Evaluate Teaching Effectiveness

Teaching effectiveness data are available from a variety of sources. These include students, peers, administrators, and others involved in the educational experience such as preceptors.

Student Ratings

Student evaluations are a necessary but insufficient source of information (Oermann, 2017). Because students are the only participants other than the teacher who are consistently present during the teaching–learning process, they have a unique perspective of the teacher’s effectiveness as an educator over time. Students can make valid and reliable interpretations about the teacher’s use of teaching methods, fairness, interest in students, and enthusiasm for the subject.

There are limitations, though, to the use of student ratings. These ratings can be affected by class size, with smaller classes tending to rate teacher effectiveness higher than larger classes (Oermann, Conklin, Rushton, & Bush, 2018). Student ratings can also be influenced by the type of course format; for example, discussion courses tend to receive higher ratings than do lecture courses. Students have a tendency to rate required and elective courses in their own field of study higher than courses they are required to take outside their majors. Lastly, it is questionable whether students can evaluate the accuracy, depth, and scope of the teacher’s knowledge because they

do not have expertise in the content to make this judgment. Characteristics such as these are best evaluated by peers.

Many colleges and universities have a standard form for student evaluation of teaching that is used in all courses across the institution. These forms generally ask students to rate the teacher's performance in areas of (a) presentation and teaching skills, (b) interactions with students as a group and individually, (c) breadth of coverage of content, and (d) assessment and grading practices. Students also may be asked to provide a rating of the overall quality of the faculty member's teaching in the course, the extent of their own learning in the course, and the workload and difficulty of the course. Table 18.2 lists typical areas that are assessed by students on these forms.

TABLE 18.2 Typical Areas Assessed on Student Evaluation of Teaching Forms

Presentation or Teaching Skills
Organized course well Gave clear explanations Used examples, illustrations, and other methods to promote understanding of content Was well prepared for class Was enthusiastic about content and teaching Stimulated students' interest in subject Motivated students to do best work Used learning activities, readings, and assignments that facilitated understanding of course content Had realistic appreciation of time and effort for students to complete assignments and course work
Interactions With Students Individually and in Groups
Encouraged student participation and discussion Showed respect for students' views and opinions Was readily available to students (e.g., questions after class, by email, by appointment)
Assessment and Grading Practices
Communicated student responsibilities clearly Explained course assignments, assessment methods, and grading procedures Was fair in assessment and grading Provided prompt and valuable feedback
Overall Course Evaluation
Course difficulty (e.g., rated on scale of <i>too difficult</i> to <i>too elementary</i>) Workload in course (e.g., rated on scale of <i>too heavy</i> to <i>too light</i>) Course pace (e.g., rated on scale of <i>too fast</i> to <i>too slow</i>) Extent of learning in course (e.g., rated on scale of <i>a great deal</i> to <i>nothing new</i>) Overall course rating (e.g., rated on scale of <i>excellent</i> to <i>poor</i>)
Overall Teacher Evaluation
Overall quality of faculty member's teaching (e.g., rated on scale of <i>excellent</i> to <i>poor</i>)

TABLE 18.3 Sample Questions for Evaluating Effectiveness of Clinical Teachers

Clinical Teacher Evaluation
<i>Purpose:</i> These questions are intended for use in evaluating teacher effectiveness in courses with a clinical component. The questions are to be used in conjunction with the college or university student evaluation of teaching form.
Clinical Teaching Items
<p>Did the teacher:</p> <ol style="list-style-type: none"> 1. Encourage students to ask questions and express diverse views in the clinical setting? 2. Encourage application of theoretical knowledge to clinical practice? 3. Provide feedback on student strengths and weaknesses related to clinical performance? 4. Develop positive relationships with students, preceptors, others in the clinical setting? 5. Inform students of their professional responsibilities? 6. Facilitate student collaboration with members of healthcare teams? 7. Facilitate learning in the clinical setting? 8. Strive to be available in the clinical setting to assist students? <p>Was the instructor</p> <ol style="list-style-type: none"> 9. An effective clinical teacher?

These general forms, however, do not evaluate teacher behaviors important in the clinical setting. Faculty members can add questions on clinical teaching effectiveness to these general forms or can develop a separate tool for students to use in evaluating teacher performance in clinical courses. Sample questions for evaluating the effectiveness of the clinical teacher are found in Table 18.3.

Students may complete teacher evaluations in face-to-face classes, administered by someone other than the teacher and without the teacher present in the room, but in most programs course evaluations are completed online. With an online course evaluation system, it is critical that students' anonymity and confidentiality be protected and that students are assured that their identity cannot be linked with their evaluations. When students do not feel confident in the anonymity of their responses, they may choose not to complete the teacher evaluations, thereby decreasing the response rate and the reliability of the findings.

Peer Review

Another source of data for evaluating teacher effectiveness comes from peers. Peer review is a form of assessment in which instructors give feedback about teaching and learning to one another. Combined with other sources of information, such as student learning outcomes, teacher self-assessment, and student feedback, peer review of teaching can be an important component of evaluation of teaching. Peer review offers a perspective of another instructor who knows the course content and who

has experience working with students at that level of the educational program. One form of peer evaluation is observing the teacher in the classroom, clinical setting, or laboratory. Observations of teaching performance are best used for formative evaluation because there are too many variables that can influence the reliability of these observations. Observations can be influenced too easily by personal feelings, positive or negative, about the colleague.

Effective peer review of classroom teaching usually includes a preobservation meeting, the classroom observation, a postobservation debriefing, and a written summary. During the preobservation meeting, the teacher and observer discuss the teacher's plan for the class, including:

- Expected learning outcomes
- How the class fits into the overall course
- Anything in particular on which the teacher would like feedback

During the classroom observation, the observer pays particular attention to criteria such as:

- The organization of the class
- Reinforcement of major concepts
- Pacing
- Classroom atmosphere and interactions with students
- Consideration of diverse views and opinions
- Strategies to engage students in the class
- Assessment techniques

The postobservation meeting provides an opportunity to review the observation, share perspectives of what happened during the session, and set goals for the future. The observer can complete a written summary of the preobservation, observation, and postobservation processes. If the peer review is done as part of the processes for promotion and tenure, a specific form may be required for this summary. This same process can be used for observations of teaching in clinical settings and in simulation.

Peer evaluation of teaching also can be conducted for online courses. By reviewing course materials and course websites as guest users, peer evaluators of teaching in online courses can look for evidence that teachers demonstrate application of the following principles of effective instruction, such as:

- How quickly and thoroughly does the teacher respond to student questions?
- Does the teacher use group assignments, discussion forums, or peer critique of assignments to promote interaction and collaboration among students?

- Does the teacher use assignments that require the active involvement of students in their own learning?
- Does the teacher provide prompt, meaningful feedback on assignments?
- Is there evidence that students are actively engaged and spend an appropriate amount of time on course tasks?
- Does the teacher have realistically high standards for achievement of course outcomes and communicate them to students?
- Does the teacher accommodate a variety of learning styles, views, abilities, and preferences?
- Is the online course well organized, with easy-to-locate course materials and clear directions for navigating the online course?
- Is the course design and site inviting, and are graphics used appropriately?

Peers can review course syllabi, instructional materials, teaching strategies, learning activities, discussion forum questions, tests, and other documents developed for courses; materials developed for clinical courses; grants, publications, and similar materials documenting the teacher's scholarship; a teaching portfolio; and other materials. This review can be used for formative purposes, to give suggestions for further development, or for summative purposes, to make decisions about contract renewal, for annual reviews, and for tenure and promotion.

To be most effective, peer review of teaching should take place within a context of continuous improvement of the teaching–learning process. It must be supported by adequate resources for faculty development, mentoring, and modeling of effective teaching by master teachers.

Teaching Portfolio

Another approach to documenting teaching effectiveness is the use of a teaching portfolio or dossier. More than just a curriculum vitae, the portfolio is a collection of teacher-selected materials or artifacts that describe the faculty member's teaching activities in the classroom, the online environment, clinical practice, the simulation center, and other settings where the instruction took place. The materials included in the portfolio should be highly selective and organized to create a cohesive professional narrative.

Teaching portfolios may serve a specific purpose, such as for teaching improvement (formative evaluation) or promotion or tenure review (summative evaluation). A career portfolio might be assembled and used to seek a faculty position. Although portfolios can be assembled from printed materials, an electronic format is widely used.

A teaching portfolio should contain materials that illustrate the faculty member's teaching effectiveness, such as syllabi, teaching strategies, sample tests, student

TABLE 18.4 Suggested Content of a Teaching portfolio

Material From the Faculty Member
Personal philosophy of teaching Statement about teaching goals Description of teaching responsibilities (e.g., classroom instruction, online teaching, clinical instruction) List of courses taught with dates Course syllabus, sample teaching strategies, materials, assignments, online activities and discussion forum questions, tests, multimedia, and other documents from one or more courses (documents should reflect the types of courses taught, e.g., classroom, online, clinical, laboratory, seminar) An edited 5-minute videotape of a class or a segment from an online course Teaching awards and recognition of teaching effectiveness (by alumni, clinical agency personnel, others)
Material From Students
Samples of student papers, good and poor, with teacher's written comments; other products of student learning Unsolicited letters from students, alumni, and clinical agency staff who work with students addressing the faculty member's teaching effectiveness (a few that are carefully selected)
Material From Colleagues and Administrators
Peer evaluation of teaching materials
Other Documents
Self-appraisal and teaching goals (short and long term) Appendices

assignments, and online materials, to name a few. The portfolio also includes the faculty member's philosophy of teaching, which should be reflected in the documents in the portfolio. Table 18.4 lists materials typically included in a portfolio for personnel decisions, such as contract renewal and reviews for promotion and tenure.

Portfolios used for instructional improvement include these same materials, but also identify areas of teaching that need improvement and efforts to further develop teaching skills such as workshops attended. In this type of teaching portfolio, peer and administrator evaluations of teaching, a self-evaluation of strengths and weaknesses, and other documents that demonstrate areas for improvement and steps taken can be included. However, these are not appropriate for a teaching portfolio that will be used for personnel decisions.

■ Systematic Program Evaluation Plan

Developing a systematic program evaluation is important for ongoing program improvement as well as compliance with regulations and accreditation standards. Although a regulatory or accrediting body can require a nursing education program

to have a systematic evaluation plan, the plan must be useful for improving the program or it will not be appropriate. The faculty should value the plan, or it will not be implemented and changes will not be made based on its findings (Lewallen, 2017).

Evaluation Framework

The organizing framework for evaluation guides the selection of evaluation criteria to use. Many nursing education programs use accreditation standards as an organizing framework; additional criteria may be added as desired. Other programs use decision- or systems-oriented models, discussed earlier in this chapter, or other models. Again, the faculty may add additional criteria and areas to evaluate whether they are important to the nursing program.

Components

After selecting the evaluation framework, the faculty should structure a plan for data collection, analysis, and reporting. Most nursing education programs use a table structure (Table 18.5), although some use a combination of narrative and table format. The structure should include the criteria (areas for evaluation), benchmarks or expected levels of achievement, person or group responsible for data collection, assessment methods, and a data-collection time frame. Many programs also include columns for results of the analysis of data collected and actions taken (Lewallen, 2017).

Areas for Evaluation

The areas of evaluation (or criteria) include regulatory and accreditation standards and any additional areas that the nursing education program is interested in evaluating. Components that all programs need to include are criteria related to administration, faculty, students, curriculum, resources, and program outcomes (Lewallen, 2017). Examples of additional areas include student satisfaction with the quality of instruction, employer satisfaction with graduates' performance, and graduates' satisfaction with the program. Criteria may be numbered to assist with ongoing tracking of elements related to them.

Benchmarks

Benchmarks or expected levels of achievement may be set by regulatory or accrediting bodies, or they may be established by the program and be higher than those levels. For example, the state board of nursing requires a minimum first-time pass rate for the NCLEX, and accrediting bodies also set benchmarks for this area of evaluation. CCNE, as an example, specifies that the NCLEX-RN® pass rate is 80% or higher for first-time test-takers; nursing education programs, however, may set this benchmark at 90%. For other areas, such as faculty scholarly productivity, the program should

establish its own benchmarks. These benchmarks should be reasonable for the program based on factors such as past performance and the availability of resources to support goal attainment. Setting benchmarks too high results in persistent areas of deficit (Lewallen, 2017). Programs using standardized testing should specify the benchmark scores they use for making admission, progression, and graduation decisions. More than one benchmark may apply to each evaluation area. For example, if the evaluation area is “student progression through the program,” benchmarks might include grade point averages and standardized test scores.

Person or Group Responsible

The specific person or group responsible for collecting data for each evaluation area should be identified. These may include individuals such as a program director, administrative assistant, or faculty member; examples of groups include a curriculum committee or school evaluation committee. These individuals or groups may be responsible for collecting and analyzing data relevant to each area of evaluation, or they may be responsible for data collection only, with data analysis assigned to another individual or group.

Assessment Method

For each criterion or evaluation area, one or more methods of collecting data should be specified. The evaluation area and benchmarks suggest assessment methods to be used. In some cases, existing assessment instruments are used, and in other instances new instruments need to be developed. Both quantitative and qualitative methods may be used for some evaluation areas. Although there may be many ways of assessing each criterion, it is not necessary to use all possible means of collecting data (Lewallen, 2017).

Time Frame

The systematic evaluation plan should specify a time frame for data collection and analysis. The evaluation plan usually spans one academic year, and data may be collected at natural times during that year such as at the end of a semester or term or at one year after graduation.

Results

If a systematic evaluation plan table includes a column for results, the data for that year are recorded there, for example, standardized test means or first-time NCLEX pass rate. The data source should be indicated, such as the minutes of a certain committee meeting (with the date), so that auditors may track and confirm the results. It is helpful to label discussions of program evaluation results in the minutes with the criterion number so that they can be easily tracked and faculty members will be aware of the ongoing nature of program evaluation (Lewallen, 2017).

Actions Taken

The last column of the evaluation plan (Table 18.5) may include decisions made and actions taken based on evaluation results. If the benchmark has been met for a criterion, this column simply indicates maintenance of the current processes. If a benchmark has not been met, an action plan with specific dates for reassessment should be included.

Content of Evaluation Areas or Criteria

As previously indicated, despite the evaluation framework or model used, each systematic evaluation plan should include certain components. The content of each of those components will be briefly described here.

Administration

The nursing education program should demonstrate how its mission, goals, and desired outcomes are aligned with those of the parent institution. This can be demonstrated by use of a table comparing those elements between the nursing education program and the institution. Budgetary support for the nursing education program in relation to the work of the program and in comparison with similar units in the institution should be included. Documentation of faculty and student participation in the work of the program and the institution also is important. It is important to note student attendance at meetings in the minutes, and even when students cannot be present, the minutes can reflect consideration of student requests and feedback (Lewallen, 2017). An important area to include in this component is the accuracy of public information about the program, including accreditation status, admission and progression policies, grievance procedures, curriculum plans, and graduation requirements. All such information must be consistent in all places in which it appears, including print and online locations. One individual may be assigned to monitor this information at least annually, and preferably more frequently, especially in electronic sources.

Faculty

The goal of this component is to demonstrate that faculty members remain qualified for their positions and that they meet the requirements of their positions. Data include educational preparation, licensure, certification, and continuing education of faculty members and preceptors (if used). Data that demonstrate faculty members meet the expectations of their positions include teaching and advisement workloads; scholarly productivity; and service to the institution, community, and profession.

TABLE 18.5 Sample Format for a Systematic Evaluation Plan

AREAS FOR EVALUATION OR CRITERIA	BENCHMARK OR EXPECTED LEVEL OF ACHIEVEMENT	PERSON OR GROUP RESPONSIBLE	ASSESSMENT METHODS	TIME FRAME	RESULTS	ACTIONS TAKEN
<i>Administration</i>						
A.1						
A.2						
A.3						
<i>Faculty</i>						
F.1						
F.2						
F.3						
<i>Students</i>						
S.1						
S.2						
S.3						
<i>Curriculum</i>						
C.1						
C.2						
C.3						
<i>Resources</i>						
R.1						
R.2						
R.3						
<i>Program Outcomes</i>						
O.1						
O.2						
O.3						

Students

In this component, areas to evaluate include availability and use of student services, maintenance of student records, and communication of policies affecting students, including grievance policies (Lewallen, 2017). It is important to monitor

the number of student complaints about the program and how they are resolved. These data also are required by the regional accrediting bodies (for institutional accreditation).

Curriculum

Areas to evaluate in this component include inclusion of professional standards in the curriculum, evaluation of courses and the overall curriculum, and how students' clinical placements contribute to their achievement of program and course outcomes. Data collection should include minutes of all meetings in which the curriculum is discussed, student and faculty course and clinical site evaluations, student performance on standardized tests, and employers' satisfaction with graduates' performance. Meeting minutes are an essential source of information about curriculum decisions made on the basis of program evaluation results (Lewallen, 2017).

Resources

In this component, areas to assess include the adequacy of fiscal, physical, and personnel resources. Data to monitor include the ratio of students to academic advisors; availability of information technology support; number of classrooms and their capacity; faculty office space; numbers and availability of staff members; availability of tutoring; simulation lab capacity and use; tutors and counseling resources; clinical site availability and effectiveness; and funding for faculty and staff positions, faculty development, and equipment purchases, among other data.

Program Outcomes

Effective evaluation of program outcomes depends on internal and external means of assessment. External means include first-time pass rates on NCLEX and certification exams, scores on standardized tests, and surveys of employers. Internal means include assessment of student work such as performance on a comprehensive exam, dissertation or scholarly project, or course assignments; graduation and attrition rates and length of time for program completion; and graduates' satisfaction with the program, perception of their achievement of program goals, readiness for the roles for which they were prepared, employment rates, and pursuit of higher degrees. It is important to note that collecting data from employers and alumni typically results in low response rates that affect the reliability of results; a combination of qualitative and quantitative approaches (such as focus groups and social media) often are more effective in collecting data from these two groups than surveys alone. Evaluation plans are generally presented in an electronic format. Opsahl and Horton-Deutsch (2019) developed a dashboard to use to communicate outcomes to stakeholders.

■ Summary

Accreditation is a means of ensuring that institutions of higher education and their programs meet standards of quality. In this way, accreditation protects the public. Accreditation involves internally reviewing and assessing one's own programs on a continuous basis to identify areas for revisions and participating in an external review, including a site visit, to verify the standards are met. There are two types of accreditation: institutional (at the university or college level) and programmatic (at the program or discipline level).

There are three accrediting bodies for nursing education programs: ACEN, CCNE, and CNEA. Two of these organizations (ACEN and CNEA) accredit all levels of nursing education programs: practical, diploma, associate, baccalaureate, masters, and DNP. CCNE accredits baccalaureate, masters, DNP, postgraduate APRN certificate programs, and entry-to-practice residency programs. The chapter also presented standards for distance education programs, which are applicable to nursing.

Program evaluation is the process of judging the worth or value of an educational program for the purposes of making decisions about the program or to provide evidence of its effectiveness in response to demands for accountability. A number of models can be used for program evaluation. Accreditation models are designed to determine whether a program meets external standards of quality. Decision-oriented models usually focus on internal standards of quality, value, and efficacy to provide information for making decisions about the program. Systems-oriented approaches consider the inputs, processes or operations, and outputs or outcomes of an educational program. The process of program evaluation assists various audiences or stakeholders of an educational program in judging its worth. Audiences or stakeholders are individuals and groups who are affected directly or indirectly by the decisions made, such as students, teachers, employers, clinical agencies, and the public. An important decision when planning a program evaluation is whether to use external or internal evaluators, or both.

One area of program evaluation involves determining the quality of teaching in the classroom and clinical setting and other dimensions of the teacher's role, depending on the goals and mission of the nursing program. The literature suggests five characteristics and qualities of effective teaching in nursing: (a) knowledge in the area in which teaching, (b) clinical competence, (c) teaching skill, (d) interpersonal relationships with students, and (e) personal characteristics. Teaching effectiveness data are available from a variety of sources, including students, faculty peers, and administrators. The use of a teaching portfolio as a way to document teaching effectiveness is another approach that allows the teacher to select and comment on items that reflect implementation of a personal philosophy of teaching.

The chapter also discussed development of a systematic evaluation plan. After selecting an evaluation framework, a plan for data collection, analysis, and reporting should be structured, usually in table format. Components of the plan include criteria (areas for evaluation), benchmarks or expected levels of achievement, person or group responsible for data collection, assessment methods, a data-collection time frame, and results of analysis of data collected and actions taken. The areas for evaluation should include administration, faculty, students, resources, and program outcomes.

■ References

- Accreditation Commission for Education in Nursing. (2019a). *ACEN 2017 Accreditation manual: Section II. Policies*. Atlanta, GA: Author. Retrieved from <http://www.acennursing.net/manuals/Policies.pdf#page=60>
- Accreditation Commission for Education in Nursing. (2019b). *ACEN 2017 Accreditation manual: Section III. 2017 standards & criteria*. Atlanta, GA: Author. Retrieved from <http://www.acennursing.net/manuals/SC2017.pdf>
- Commission on Collegiate Nursing Education. (2018). *Standards for accreditation of baccalaureate and graduate degree nursing programs*. Retrieved from <https://www.aacnnursing.org/Portals/42/CCNE/PDF/Standards-Final-2018.pdf>
- Commission on Collegiate Nursing Education. (2019). *What we do*. Washington, DC: American Association of Colleges of Nursing. Retrieved from <https://www.aacnnursing.org/CCNE-Accreditation/What-We-Do>
- Commission for Nursing Education Accreditation. (2016). *Accreditation standards for nursing education programs*. Washington, DC: National League for Nursing. Retrieved from <http://www.nln.org/docs/default-source/accreditation-services/cnea-standards-final-february-201613f2bf5c78366c709642ff00005f0421.pdf?sfvrsn=4>
- Commission for Nursing Education Accreditation. (2019). *CNEA mission and values*. Washington, DC: National League for Nursing. Retrieved from [http://www.nln.org/accreditation-services/the-nln-commission-for-nursing-education-accreditation-\(cnea\)](http://www.nln.org/accreditation-services/the-nln-commission-for-nursing-education-accreditation-(cnea))
- Council of Regional Accrediting Commissions. (2011). *Interregional guidelines for the evaluation of distance education*. Retrieved from <https://www.nc-sara.org/files/docs/C-RAC%20Guidelines.pdf>
- Distance Education Accrediting Commission. (2019). *Part three: Accreditation standards. The DEAC accreditation handbook*. Washington, DC: Author. Retrieved from <https://www.deac.org/Seeking-Accreditation/The-DEAC-Accrediting-Handbook.aspx>
- Halstead, J. A. (2017). The accreditation process in nursing education. In M. H. Oermann (Ed.), *A systematic approach to assessment and evaluation of nursing programs* (pp. 79–91). Philadelphia, PA: National League for Nursing/Wolters Kluwer.
- Iwasiw, C., Andrusyszyn, M.-A., & Goldenberg, D. (2020). *Curriculum development in nursing education* (4th ed.). Burlington, MA: Jones & Bartlett Learning.
- Lewallen, L. P. (2017). Developing a systematic program evaluation plan for a school of nursing. In M. H. Oermann (Ed.), *A systematic approach to assessment and evaluation of nursing programs* (pp. 45–57). Washington, DC: National League for Nursing.

- Lovric, R., Prlic, N., Zec, D., Puseljic, S., & Zvanut, B. (2015). Students' assessment and self-assessment of nursing clinical faculty competencies: Important feedback in clinical education? *Nurse Educator*, 40, E1–E5. doi:10.1097/nne.0000000000000137
- National Council of State Boards of Nursing. (2019). *About U.S. boards of nursing*. Retrieved from <https://www.ncsbn.org/about-boards-of-nursing.htm>
- National League for Nursing. (2012). *The scope of practice for academic nurse educators*. Washington, DC: Author.
- National League for Nursing. (2019). *Nurse educator core competency*. Retrieved from <http://www.nln.org/professional-development-programs/competencies-for-nursing-education/nurse-educator-core-competency>
- Oermann, M. H. (2017). Student evaluations of teaching: There is more to course evaluations than student ratings. *Nurse Educator*, 42, 55–56. doi:10.1097/nne.0000000000000366
- Oermann, M. H., & Conklin, J. L. (2018). Evidence-based teaching in nursing. In M. H. Oermann, J. C. De Gagne, & B. C. Phillips (Eds.), *Teaching in nursing and role of the educator: The complete guide to best practice in teaching, evaluation, and curriculum development* (2nd ed., pp. 363–377). New York, NY: Springer Publishing Company.
- Oermann, M. H., Conklin, J. L., Rushton, S., & Bush, M. A. (2018). Student evaluations of teaching (SET): Guidelines for their use. *Nursing Forum*, 53, 280–285. doi:10.1111/nuf.12249
- Oermann, M. H., Shellenbarger, T., & Gaberson, K. B. (2018). *Clinical teaching strategies in nursing* (5th ed.). New York, NY: Springer Publishing Company.
- Opsahl, A., & Horton-Deutsch, S. (2019). A nursing dashboard to communicate the evaluation of program outcomes. *Nurse Educator*, 44. doi:10.1097/nne.0000000000000632. [e-pub ahead of print]
- Saewert, K. J. (2017). Program evaluation perspectives and models. In M. H. Oermann (Ed.), *A systematic approach to assessment and evaluation of nursing programs* (pp. 7–18). Washington, DC: National League for Nursing.
- Stufflebeam, D. L. (2013). The CIPP evaluation model: Status, origin, development, use, and theory. In M. C. Alkin (Ed.), *Evaluation roots: A wider perspective of theorists' views and influences* (2nd ed., pp. 243–260). Los Angeles, CA: Sage.
- Valiga, T. (2017). Curriculum evaluation. In M. H. Oermann (Ed.), *A systematic approach to assessment and evaluation of nursing programs* (pp. 19–28). Washington, DC: National League for Nursing.

APPENDIX A

QUICK REFERENCE GUIDE FOR WRITING DIFFERENT TYPES OF TEST ITEMS WITH EXAMPLES

■ True–False

Guidelines for Writing

1. The true–false item should test recall of important facts and information.
2. The statement should be true or false without qualification—unconditionally true or false.
3. Avoid words such as *usually*, *sometimes*, *often*, and similar terms. Because these words typically are used more often in true statements, they give students clues to the correct response. Avoid words such as *never*, *always*, *all*, and *none*, which often signal a false response.
4. Avoid terms that indicate an infinite degree or amount, such as *large*, that can be interpreted differently by students.
5. Each item should include one idea to be tested rather than multiple ones.
6. Items should be worded precisely and clearly. Avoid long statements with different qualifiers and focus on the main idea to be tested.
7. Avoid the use of negatives, particularly in false statements.
8. With a series of true–false items, statements should be similar in length.
9. Check that the answers to true–false items are not ordered in a noticeable pattern on the test.
10. Decide how to score true–false items prior to administering them to students. In some variations of true–false items, students correct false statements; for this type, award 2 points, 1 for identifying the statement as false and 1 for correcting it. With items of this type, do not reveal the point value of each item because this would cue students that 2-point items are false.

11. Write clear and specific directions for variations of true–false items. Variations include correcting a false statement, providing a rationale for the answer (whether true or false), and multiple true–false items (incomplete statement followed by several true or false phases that complete it).

Examples

#1. For each of the following statements, select T if the statement is true and F if the statement is false:

- | | | |
|---|---|--|
| T | F | Rocky Mountain spotted fever is caused by <i>Rickettsia rickettsii</i> , which is carried by infected ticks. (T) |
| T | F | S2 heart sound is caused by opening of the semilunar valves. (F) |
| T | F | The diaphragm of the stethoscope is used for auscultating S3 and S4 and low-pitched tones. (F) |
| T | F | A visual acuity of 20/30 indicates the patient can see at 20 ft what a person with normal vision can see at 30 ft. (T) |

#2. If the statement is true, select T and do no more. If the statement is false, select F, underline the word or phrase that makes it false, and write in the blank the word or phrase that would make it true.

- | | | |
|---|----------|--|
| T | <u>F</u> | S2 heart sound is caused by <u>opening</u> of the semilunar valves.
closure _____ |
|---|----------|--|

#3. If the statement is true, select T and do no more. If the statement is false, select F and select the *correct* answer from the list that follows the item.

- | | | |
|---|----------|--|
| T | <u>F</u> | A split S2 is best heard in the <u>mitral</u> area.
aortic area, <u>pulmonic area</u> , tricuspid area, Erb's point |
|---|----------|--|

■ Matching

Guidelines for Writing

1. Develop a matching exercise around homogeneous content. All of the premises and responses to be matched should relate to that content, for example, all laboratory tests and values, all terms and definitions. This is the most important principle in writing a matching exercise.
2. Include an unequal number of premises and responses to avoid giving a clue to the final match. Typically there are more responses than premises.

3. Use short lists of premises and responses.
4. For matching exercises with a large number of responses, develop two separate matching exercises.
5. Directions for the matching exercises should be clear and state explicitly the basis for matching the premises and responses. Even if the basis for matching seems self-evident, the directions should include the rationale for matching the columns.
6. Directions should specify whether each response may be used once, more than once, or not at all. The directions must be clear about the selection of responses.
7. Place the longer premises on the left and shorter responses on the right. This enables the students to read the longer statement first, then search on the right for the correct response, which often is a single word or a few words.
8. The order in which the premises and responses are listed should be alphabetical, numerical, or in some other logical order. If the lists have another logical order, however, such as dates and sequential steps of a procedure, then they should be organized in that order. Numbers, quantities, and similar types of items should be arranged in decreasing or increasing order.
9. The entire matching exercise should be placed on the same page and in the same box if online.

Examples

#4. Directions: For each cranial nerve in Column A, select the proper function in Column B. Use each letter only once or not at all.

Column A

- _____ 1. Abducens
 _____ 2. Oculomotor
 _____ 3. Olfactory
 _____ 4. Trochlear
 _____ 5. Vestibulocochlear

Column B

- a. Pupil response
 b. Hearing and balance
 c. Outward eye movement
 d. Downward, outward, and inward eye movements
 e. Smells
 f. Taste
 g. Tongue muscle movements

Answers to #4:

Column A

- c___ 1. Abducens
a___ 2. Oculomotor
e___ 3. Olfactory
d___ 4. Trochlear
b___ 5. Vestibulocochlear

Column B

- a. Pupil response
b. Hearing and balance
c. Outward eye movement
d. Downward, outward, and inward eye movements
e. Smells
f. Taste
g. Tongue muscle movements

#5. Directions: For each brand name in Column A, select the generic name in Column B. Use each letter only once or not at all.

Column A

- _____ 1. Apresoline
_____ 2. Cardizem
_____ 3. Lopressor
_____ 4. Norvasc
_____ 5. Tenormin

Column B

- a. Amlodipine
b. Atenolol
c. Diltiazem
d. Hydralazine
e. Metoprolol
f. Propranolol

Answers to #5:

Column A

- d___ 1. Apresoline
c___ 2. Cardizem
e___ 3. Lopressor
a___ 4. Norvasc
b___ 5. Tenormin

Column B

- a. Amlodipine
b. Atenolol
c. Diltiazem
d. Hydralazine
e. Metoprolol
f. Propranolol

■ Multiple Choice

Guidelines for Writing: *Stem*

1. The stem should present clearly and explicitly the problem to be solved. The student should not have to read the alternatives to understand the question or the incomplete statement.
2. Although the stem should be clear and explicit, it should not contain extraneous information unless the item is developed for the purpose of identifying significant versus insignificant data.
3. Avoid inserting information in the stem for instructional purposes.
4. If words need to be repeated in each alternative to complete the statement, shift them to the stem.
5. Do not include key words in the stem that would clue the student to the correct answer.
6. Avoid the use of negatively stated stems, including words such as *no*, *not*, and *except*.
7. The stem and alternatives that follow should be consistent grammatically.
8. Avoid ending stems with *a* or *an* because these often provide grammatical clues as to the option to select.
9. If the stem is a statement completed by the alternatives, begin each alternative with a lowercase letter and place a period after it because it forms a sentence with the stem. At the end of the stem, use a comma or colon as appropriate.
10. Each multiple-choice item should be independent of the others.
11. Write the stem so that the alternatives are placed at the end of the incomplete statement.

Guidelines for Writing: *Alternatives*

1. The alternatives should be similar in length, detail, and complexity, and should have the same number of parts.
2. The alternatives should be consistent grammatically.
3. The alternatives should sample the same domain, for instance, all symptoms, all diagnostic tests, all nursing interventions, varying treatments, and so forth.
4. Avoid including opposite responses among the options. This is often a clue to choose between the opposites and not consider the others.

5. Arrange the options in a logical or meaningful order (alphabetical, numerical, or chronological).
6. Options with numbers, quantities, and other numerical values should be listed sequentially, either increasing or decreasing in value, and the values should not overlap.
7. Each option should be placed on a separate line for ease of student reading.
8. Use the option of *call for assistance* and *notify the physician* sparingly.

Guidelines for Writing: *Correct or Best Answer*

1. Review the alternatives carefully to ensure that there is only one correct response.
2. Review terminology included in the stem carefully to avoid giving a clue to the correct answer.
3. The correct answer should be randomly assigned to a position among the alternatives to avoid favoring a particular response choice.
4. The answers should not reflect the opinion of the teacher but instead should be the ones with which experts agree or are the most probable responses.

Guidelines for Writing: *Distractors*

1. The distractors should be consistent grammatically and should be similar in length, detail, and complexity with each other and the correct answer.
2. The distractors should sample the same content area as the correct answer.
3. Avoid using *all of the above* and *none of the above* in a multiple-choice item.
4. Omit terms such as *always*, *never*, *sometimes*, *occasionally*, and similar ones from the distractors.
5. Avoid using distractors that are essentially the same.

Examples

#6. Your patient is 50 years old. You measure her height as 63 in. and weight as 148 lbs. Using a 22- to 25-gauge needle, which is the appropriate site and needle size to use to administer an intramuscular immunization to this patient?

- a. Anterolateral site, 1-in. needle size
- b. Anterolateral site, 5/8-in. needle size
- c. Deltoid or anterolateral site, 1- to 1¼ in. needle size
- d. Deltoid site, 1- to 1½ in. needle size¹

#7. A 14-year-old boy is brought to the health clinic by his mother for a sports physical. He had his last vaccines when he was 6 years old. Which of the following vaccines are recommended by the Centers for Disease Control and Prevention for this patient?

- a. DT and MCV4 vaccines
- b. Td and HPV vaccines
- c. Tdap and influenza vaccines
- d. Tdap, MCV4, and the HPV vaccines¹

#8. In morning report, you are told that your patient will be started on a selective serotonin reuptake inhibitor (SSRI). Which of these medications is an SSRI?

- a. Fluoxetine (Prozac)
- b. Lithium carbonate (Lithobid)
- c. Paroxetine (Paxil)¹
- d. Risperidone (Risperdal)

#9. Your patient had a transurethral resection of the prostate (TURP). You should recognize the need for further teaching when the patient states:

- a. “I’ll have a catheter after surgery.”
- b. “I’ll need to stay in the hospital for at least 3 to 5 days.”¹
- c. “I need to be careful not to do heavy lifting right after surgery.”
- d. “The TURP procedure was done because of my enlarged prostate.”

¹Correct answer.

Exhibit format:

#10. Your patient is an elderly man with a long history of smoking. In report, you are told he was admitted with a severe cough, blood-tinged sputum, and shortness of breath.

The medical record shows:

Weight 118 lb

Heart rate (HR): 120 beats per minute (bpm)

Respiratory rate (RR): labored, 30 bpm

Blood pressure (BP): 110/50 mmHg

Oxygen saturation: 74% on room air

Your priority concern should be his:

- a. BP.
- b. HR.
- c. oxygen saturation.¹
- d. weight.

Ordered-response format:

#11. Your patient has severe abdominal pain, bloody diarrhea, and rectal irritation. He is ordered to have nothing by mouth (NPO) and to have morphine intravenously (IV) for pain control via a pump. His white blood cells are 18,000, hemoglobin is 10.2, and hematocrit is 28. The tentative diagnosis is ruptured diverticula.

Place these nursing interventions in their order of importance from 1 to 3:

- ___ Hang 0.9% normal saline IV solution.
- ___ Place a urinary catheter.
- ___ Request an order for antibiotic therapy.

Answers to #11:

- 1 Hang 0.9% normal saline IV solution.
- 3 Place a urinary catheter.
- 2 Request an order for antibiotic therapy.

■ Multiple Response

1. The combination of alternatives should be plausible. Options should be logically combined rather than grouped randomly.
2. The alternatives should be used a similar number of times in the combinations.
3. The responses should be listed in a logical order, for instance, alphabetically or sequentially, for ease in reviewing.

Examples

#12. Select all that apply.

The following statements about herpes zoster are true:

- ✓ 1. Herpes zoster is infectious until it is crusty.
- 2. Herpes zoster is often preceded by a pattern of papulovesicles along a dermatome.
- 3. Removing the scabs decreases scarring.
- ✓ 4. Scarring may occur with herpes zoster.
- ✓ 5. The onset of herpes zoster can be preceded by itching, tingling, or pain several days before eruption.
- 6. There is a cure for herpes zoster.

#13. Select the best combination of responses:

Your patient has been placed on phenelzine (Nardil) and should avoid which of these foods?

- a. Aged cheese
 - b. Apples
 - c. Bananas and raisins
 - d. Pepperoni
 - e. Red wine
1. a, b, c
 2. a, b, d
 3. a, c, d, e¹
 4. d, e

■ Short Answer (Fill-in-the-Blank)

Guidelines for Writing

1. Questions and statements should not be taken verbatim from textbooks, other readings, or handouts students received.
2. Phrase the item so that a unique word, series of words, or number must be supplied to complete it.
3. Write questions that are specific and can be answered in a few words, phrases, or short sentences.
4. Before writing the item, think of the correct answer first and then write a question or statement for that answer.
5. Fill-in-the-blank items requiring calculations and that solve mathematical-type problems should include in the statement the type of answer and degree of specificity desired.
6. For a statement with a key word or words missing, place the blank at or near the end of the statement.
7. When students need to write longer answers, provide sufficient space.
8. Even though a blank space is placed at the end of the statement, the teacher may direct the student to record one-word answers in blanks arranged in a column to the left or right of the items, thereby facilitating scoring.

Examples

#14. The normal respiratory rate for a newborn is how many breaths per minute?
_____ to _____

Answer: 30 to 60 breaths per minute

Calculation format

#15. Your order is to give terbutaline (Brethine) 0.25 mg subcutaneous. You have terbutaline 1 mg in 1 mL on hand. How much should you give? _____

Answer: 0.25 mg

#16. The newborn weighs 8 lb 6 oz. How many kilograms is that? _____

Answer: 3.8 kg

Hot Spot format

#17. A 6-month-old infant is admitted with symptoms of respiratory distress syndrome. Draw a line to the area where retractions would be assessed and label each line using the supplied terms:

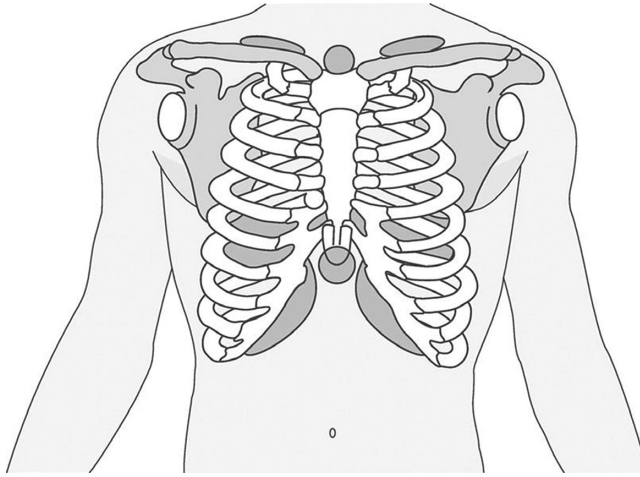
Supraclavicular

Suprasternal

Intercostal

Substernal

Subcostal



■ Essay

Guidelines for Writing

1. Develop essay items that require synthesis of the content.
2. Phrase items clearly.
3. Prepare students for essay tests.
4. Tell students about apportioning their time to allow sufficient time for answering each essay item.
5. Score essay items that deal with the analysis of issues according to the rationale that students develop rather than the position they take on the issue.
6. Avoid the use of optional items and student choice of items to answer.
7. In the process of developing the item, write an ideal answer to it.
8. If possible, have a colleague review the item and explain how he or she would respond to it.

Guidelines for Scoring Essay Items

1. Identify the method of scoring to be used prior to the testing situation and inform the students of it.
2. Specify an ideal answer in advance.
3. If using a scoring rubric, discuss it with the students ahead of time so that they are aware of how their essay responses will be judged.
4. Read a random sample of papers to get a sense of how the students approached the items and an idea of the overall quality of the answers.
5. Score all the answers to one item at a time.
6. Read each answer twice before scoring.
7. Read papers in random order.
8. Use the same scoring system for all papers.
9. Read essay answers and other written assignments anonymously.
10. Cover the scores of the previous answers to avoid being biased about the student's ability.
11. For important decisions or if unsure about the scoring, have a colleague read and score the answers to improve reliability.
12. Adopt a policy on writing (e.g., spelling, grammar).

Examples

1. Compare metabolic and respiratory acidosis. Include the following in your response: definitions, risk factors, symptoms, diagnostic tests, and interventions.
2. Your elderly patient, who lives with her grandson, tells you that her grandson locks her in the bathroom when he goes to work. The patient is frail, is poorly groomed, and has an odor. What additional data would you collect and why? Develop an action plan with a rationale.

APPENDIX B

TESTING RESOURCES FOR NURSE EDUCATORS

<p>Appropriate Use of High-Stakes Testing in Our Nation's Schools</p> <p>https://www.apa.org/pubs/info/brochures/testing</p>	<p>Principles for high-stakes testing that is sound with results used appropriately</p>
<p>Code of Fair Testing Practices in Education</p> <p>https://www.apa.org/science/programs/testing/fair-testing.pdf</p> <p>[Provided in Appendix C]</p>	<p>Guidelines for developing and using tests that are fair to all students and other test-takers regardless of their "age, gender, disability, race, ethnicity, national origin, religion, sexual orientation, linguistic background, or other personal characteristics" (p. 2)</p>
<p>Code of Professional Responsibilities in Educational Measurement</p> <p>http://www.ncme.org/resources/library/professional-responsibilities</p>	<p>Guidelines for all individuals involved in educational assessment to ensure that it is conducted in a professionally responsible manner; code includes eight major areas of assessment and expectations of individuals involved in assessment</p>
<p>National Council of State Boards of Nursing</p> <p>https://www.ncsbn.org/nclex.htm</p>	<p>Describes development and testing of NCLEX® (National Council Licensure Examination), includes a video and other resources on computer adaptive testing, provides NCLEX test plans, and has other resources for nurse educators on developing test items</p>
<p>National League for Nursing Fair Testing Guidelines for Nursing Education</p> <p>http://www.nln.org/docs/default-source/default-document-library/fairtestingguidelines.pdf?sfvrsn=2</p> <p>[Provided in Appendix D]</p>	<p>Guidelines to ensure ethical and fair testing practices in nursing education; supports the view that tests and other evaluation measures are used not only to evaluate achievement but also for learning, improving teaching, and evaluating programs</p>

(continued)

(continued)

Rights and Responsibilities of Test Takers: Guidelines and Expectations https://www.apa.org/science/programs/testing/rights	Describes rights of test-takers in the testing process
Standards for Teacher Competence in Educational Assessment of Students https://buros.org/standards-teacher-competence-educational-assessment-students [Provided in Appendix E]	Identifies seven standards related to educational assessment in which teachers should be competent, for example, be skilled in choosing appropriate assessment; in administering, scoring, and interpreting assessment results; and others
Testing and Assessment https://www.apa.org/science/programs/testing	Reports from the American Psychological Association and links related to resources on testing

APPENDIX C

CODE OF FAIR TESTING PRACTICES IN EDUCATION

■ Prepared by the Joint Committee on Testing Practices

The Code of Fair Testing Practices in Education (*Code*) is a guide for professionals in fulfilling their obligation to provide and use tests that are fair to all test-takers regardless of age, gender, disability, race, ethnicity, national origin, religion, sexual orientation, linguistic background, or other personal characteristics. Fairness is a primary consideration in all aspects of testing. Careful standardization of tests and administration conditions helps to ensure that all test-takers are given a comparable opportunity to demonstrate what they know and how they can perform in the area being tested. Fairness implies that every test-taker has the opportunity to prepare for the test and is informed about the general nature and content of the test, as appropriate to the purpose of the test. Fairness also extends to the accurate reporting of individual and group test results. Fairness is not an isolated concept, but must be considered in all aspects of the testing process.

The *Code* applies broadly to testing in education (admissions, educational assessment, educational diagnosis, and student placement) regardless of the mode of presentation, so it is relevant to conventional paper-and-pencil tests, computer-based tests, and performance tests. It is not designed to cover employment testing, licensure or certification testing, or other types of testing outside the field of education. The *Code* is directed primarily at professionally developed tests used in formally administered testing programs. Although the *Code* is not intended to cover tests made by teachers for use in their own classrooms, teachers are encouraged to use the guidelines to help improve their testing practices.

The *Code* addresses the roles of test developers and test users separately. Test developers are people and organizations who construct tests, as well as those who set policies for testing programs. Test users are people and agencies who select tests, administer tests, commission test development services, or make decisions on the basis of test scores. Test-developer and test-user roles may overlap, for example, when a state or local education agency commissions test development services, sets

policies that control the test development process, and makes decisions on the basis of the test scores.

Many of the statements in the *Code* refer to the selection and use of existing tests. When a new test is developed, when an existing test is modified, or when the administration of a test is modified, the *Code* is intended to provide guidance for this process.

The *Code* is not intended to be mandatory, exhaustive, or definitive, and may not be applicable to every situation. Instead, the *Code* is intended to be aspirational and is not intended to take precedence over the judgment of those who have competence in the subjects addressed.

The *Code* provides guidance separately for test developers and test users in four critical areas:

- A. Developing and Selecting Appropriate Tests
- B. Administering and Scoring Tests
- C. Reporting and Interpreting Test Results
- D. Informing Test-Takers

■ **A. Developing and Selecting Appropriate Tests**

TEST DEVELOPERS	TEST USERS
Test developers should provide the information and supporting evidence that test users need to select appropriate tests.	Test users should select tests that meet the intended purpose and that are appropriate for the intended test-takers.
A-1. Provide evidence of what the test measures, the recommended uses, the intended test-takers, and the strengths and limitations of the test, including the level of precision of the test scores.	A-1. Define the purpose for testing, the content and skills to be tested, and the intended test-takers. Select and use the most appropriate test based on a thorough review of available information.
A-2. Describe how the content and skills to be tested were selected and how the tests were developed.	A-2. Review and select tests based on the appropriateness of test content, skills tested, and content coverage for the intended purpose of testing.
A-3. Communicate information about a test’s characteristics at a level of detail appropriate to the intended test users.	A-3. Review materials provided by test developers and select tests for which clear, accurate, and complete information is provided.
A-4. Provide guidance on the levels of skills, knowledge, and training necessary for appropriate review, selection, and administration of tests.	A-4. Select tests through a process that includes persons with appropriate knowledge, skills, and training.

(continued)

(continued)

TEST DEVELOPERS	TEST USERS
A-5. Provide evidence that the technical quality, including reliability and validity, of the test meets its intended purposes.	A-5. Evaluate evidence of the technical quality of the test provided by the test developer and any independent reviewers.
A-6. Provide to qualified test users representative samples of test questions or practice tests, directions, answer sheets, manuals, and score reports.	A-6. Evaluate representative samples of test questions or practice tests, directions, answer sheets, manuals, and score reports before selecting a test.
A-7. Avoid potentially offensive content or language when developing test questions and related materials.	A-7. Evaluate procedures and materials used by test developers, as well as the resulting test, to ensure that potentially offensive content or language is avoided.
A-8. Make appropriately modified forms of tests or administration procedures available for test-takers with disabilities who need special accommodations.	A-8. Select tests with appropriately modified forms or administration procedures for test-takers with disabilities who need special accommodations.
A-9. Obtain and provide evidence on the performance of test-takers of diverse subgroups, making significant efforts to obtain sample sizes that are adequate for subgroup analyses. Evaluate the evidence to ensure that differences in performance are related to the skills being assessed.	A-9. Evaluate the available evidence on the performance of test-takers of diverse subgroups. Determine to the extent feasible which performance differences may have been caused by factors unrelated to the skills being assessed.

■ B. Administering and Scoring Tests

TEST DEVELOPERS	TEST USERS
Test developers should explain how to administer and score tests correctly and fairly.	Test users should administer and score tests correctly and fairly.
B-1. Provide clear descriptions of detailed procedures for administering tests in a standardized manner.	B-1. Follow established procedures for administering tests in a standardized manner.
B-2. Provide guidelines on reasonable procedures for assessing persons with disabilities who need special accommodations or those with diverse linguistic backgrounds.	B-2. Provide and document appropriate procedures for test-takers with disabilities who need special accommodations or those with diverse linguistic backgrounds. Some accommodations may be required by law or regulation.

(continued)

(continued)

TEST DEVELOPERS	TEST USERS
B-3. Provide information to test-takers or test users on test question formats and procedures for answering test questions, including information on the use of any needed materials and equipment.	B-3. Provide test-takers with an opportunity to become familiar with test question formats and any materials or equipment that may be used during testing.
B-4. Establish and implement procedures to ensure the security of testing materials during all phases of test development, administration, scoring, and reporting.	B-4. Protect the security of test materials, including respecting copyrights and eliminating opportunities for test-takers to obtain scores by fraudulent means.
B-5. Provide procedures, materials, and guidelines for scoring the tests, and for monitoring the accuracy of the scoring process. If scoring the test is the responsibility of the test developer, provide adequate training for scorers.	B-5. If test scoring is the responsibility of the test user, provide adequate training to scorers and ensure and monitor the accuracy of the scoring process.
B-6. Correct errors that affect the interpretation of the scores and communicate the corrected results promptly.	B-6. Correct errors that affect the interpretation of the scores and communicate the corrected results promptly.
B-7. Develop and implement procedures for ensuring the confidentiality of scores.	B-7. Develop and implement procedures for ensuring the confidentiality of scores.

■ **C. Reporting and Interpreting Test Results**

TEST DEVELOPERS	TEST USERS
Test developers should report test results accurately and provide information to help test users interpret test results correctly.	Test users should report and interpret test results accurately and clearly.
C-1. Provide information to support recommended interpretations of the results, including the nature of the content, norms or comparison groups, and other technical evidence. Advise test users of the benefits and limitations of test results and their interpretation. Warn against assigning greater precision than is warranted.	C-1. Interpret the meaning of the test results, taking into account the nature of the content, norms or comparison groups, other technical evidence, and benefits and limitations of test results.
C-2. Provide guidance regarding the interpretations of results for tests administered with modifications. Inform test users of potential problems in interpreting test results when tests or test administration procedures are modified.	C-2. Interpret test results from modified test or test administration procedures in view of the impact those modifications may have had on test results.

(continued)

TEST DEVELOPERS	TEST USERS
C-3. Specify appropriate uses of test results and warn test users of potential misuses.	C-3. Avoid using tests for purposes other than those recommended by the test developer unless there is evidence to support the intended use or interpretation.
C-4. When test developers set standards, provide the rationale, procedures, and evidence for setting performance standards or passing scores. Avoid using stigmatizing labels.	C-4. Review the procedures for setting performance standards or passing scores. Avoid using stigmatizing labels.
C-5. Encourage test users to base decisions about test-takers on multiple sources of appropriate information, not on a single test score.	C-5. Avoid using a single test score as the sole determinant of decisions about test-takers. Interpret test scores in conjunction with other information about individuals.
C-6. Provide information to enable test users to accurately interpret and report test results for groups of test-takers, including information about who were and who were not included in the different groups being compared, and information about factors that might influence the interpretation of results.	C-6. State the intended interpretation and use of test results for groups of test-takers. Avoid grouping test results for purposes not specifically recommended by the test developer unless evidence is obtained to support the intended use. Report procedures that were followed in determining who were and who were not included in the groups being compared and describe factors that might influence the interpretation of results.
C-7. Provide test results in a timely fashion and in a manner that is understood by the test-taker.	C-7. Communicate test results in a timely fashion and in a manner that is understood by the test-taker.
C-8. Provide guidance to test users about how to monitor the extent to which the test is fulfilling its intended purposes.	C-8. Develop and implement procedures for monitoring test users, including consistency with the intended purposes of the test.

■ D. Informing Test-Takers

Test developers or test users should inform test-takers about the nature of the test, test-taker rights and responsibilities, the appropriate use of scores, and procedures for resolving challenges to scores.

D-1. Inform test-takers in advance of the test administration about the coverage of the test, the types of question formats, the directions, and appropriate test-taking strategies. Make such information available to all test-takers.

D-2. When a test is optional, provide test-takers or their parents/guardians with information to help them judge whether a test should be taken—including indications of any consequences that may result from not taking the test (e.g., not being

eligible to compete for a particular scholarship)—and whether there is an available alternative to the test.

D-3. Provide test-takers or their parents/guardians with information about rights test-takers may have to obtain copies of tests and completed answer sheets, to retake tests, to have tests rescored, or to have scores declared invalid.

D-4. Provide test-takers or their parents/guardians with information about responsibilities test-takers have, such as being aware of the intended purpose and uses of the test, performing at capacity, following directions, and not disclosing test items or interfering with other test-takers.

D-5. Inform test-takers or their parents/guardians how long scores will be kept on file and indicate to whom, under what circumstances, and in what manner test scores and related information will or will not be released. Protect test scores from unauthorized release and access.

D-6. Describe procedures for investigating and resolving circumstances that might result in canceling or withholding scores, such as failure to adhere to specified testing procedures.

D-7. Describe procedures that test-takers, parents/guardians, and other interested parties may use to obtain more information about the test, register complaints, and have problems resolved.

Under some circumstances, test developers have direct communication with the test-takers and/or control of the tests, testing process, and test results. In other circumstances, the test users have these responsibilities.

The *Code* is intended to be consistent with the relevant parts of the Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999). The *Code* is not meant to add new principles over and above those in the Standards or to change their meaning. Rather, the *Code* is intended to represent the spirit of selected portions of the Standards in a way that is relevant and meaningful to developers and users of tests, as well as to test-takers and/or their parents or guardians. States, districts, schools, organizations, and individual professionals are encouraged to commit themselves to fairness in testing and safeguarding the rights of test-takers. The *Code* is intended to assist in carrying out such commitments.

The *Code* has been prepared by the Joint Committee on Testing Practices, a cooperative effort among several professional organizations. The aim of the Joint Committee is to act, in the public interest, to advance the quality of testing practices. Members of the Joint Committee include the American Counseling Association (ACA), the AERA, the APA, the American Speech–Language–Hearing Association (ASHA), the National Association of School Psychologists (NASP), the National Association of Test Directors (NATD), and the NCME.

Note: The membership of the Working Group that developed the *Code of Fair Testing Practices in Education* and of the Joint Committee on Testing Practices that guided the Working Group is as follows:

Peter Behuniak, PhD

Lloyd Bond, PhD

Gwyneth M. Boodoo, PhD

Wayne Camara, PhD

Ray Fenton, PhD

John J. Fremer, PhD (Cochair)

Sharon M. Goldsmith, PhD

Bert F. Green, PhD

William G. Harris, PhD

Janet E. Helms, PhD

Stephanie H. McConaughy, PhD

Julie P. Noble, PhD

Wayne M. Patience, PhD

Carole L. Perlman, PhD

Douglas K. Smith, PhD (deceased)

Janet E. Wall, EdD (Cochair)

Pat Nellor Wickwire, PhD

Mary Yakimowski, PhD

Lara Frumkin, PhD, of the APA served as staff liaison.

APPENDIX D

NATIONAL LEAGUE FOR NURSING FAIR TESTING GUIDELINES FOR NURSING EDUCATION

Developed by the National League for Nursing (NLN) Presidential Task Force on High-Stakes Testing, the Fair Testing Guidelines for Nursing Education are based on the League's core values of caring, integrity, diversity, and excellence, and on widely accepted testing principles. Fair, in this context, means that all test-takers are given comparable opportunities to demonstrate what they know and are able to do in the learning area being tested (Code of Fair Testing Practices in Education, 2004).

These guidelines have been formulated within the context of an overall need for testing. We acknowledge that faculty are fully committed to assessing students' abilities and to assuring that they are competent to practice nursing. Faculty are also cognizant that current approaches to learning assessment are limited and imperfect.

The NLN supports the belief that tests and evaluative measures are used not only to evaluate student achievement but, as importantly, to support student learning, and evaluate and improve teaching and program effectiveness. Within this framework, the standards for testing in high-stakes situations are consistent with general practices for ethical and fair testing practices.

The NLN Fair Testing Guidelines for Nursing Education value students' perspectives and backgrounds and acknowledge the role of faculty in their implementation.

I. General Guidelines

- A. Faculty have an ethical obligation to ensure that both tests and the decisions based on tests are valid, supported by solid evidence, consistent across their programs, and fair to all test-takers regardless of age, gender, disability, race, ethnicity, national origin, religion, sexual orientation, linguistic background, testing style and ability, or other personal characteristics.
- B. Faculty have the responsibility to assess students' abilities and assure that they are competent to practice nursing, while recognizing that current approaches to learning assessment are limited and imperfect.

- C. Multiple sources of evidence are required to evaluate basic nursing competence. Multiple approaches for assessment of knowledge and clinical abilities are particularly critical when high-stakes decisions (such as progression or graduation) are based on the assessment.
 - D. Tests and other evaluative measures are used not only to evaluate students' achievements, but, as importantly, to support student learning, improve teaching, and guide program improvements.
 - E. Standardized tests must have comprehensive testing, administration, and evaluation information readily available to faculty before they administer, grade, distribute results, or write related policies for test results. Faculty have the responsibility to review and incorporate these materials in communications to students about standardized testing and its consequences.
 - F. Faculty and schools of nursing have an ethical obligation to protect every student's right to privacy by maintaining appropriate confidentiality related to the testing process and to test results.
- II. Test Development and Implementation
- A. Selecting Appropriate Tests
 - 1. Standardized tests must show evidence of reliability, content and predictive validity, and evidence of fairness and equity as shown by test performance across test-taking subgroups based on culture, race, or gender.
 - 2. Tests should be appropriate to their purpose and have good technical quality.
 - 3. Tests should be screened for offensive content or scenarios.
 - 4. Tests should be reviewed regularly for content accuracy and relevance to practice.
 - 5. Test vendors should provide technical manuals that provide information on the test's blueprint, test development procedures, psychometric testing, and norms.
 - B. Informing Test-Takers
 - 1. Students should be notified as early as possible about the nature and content of the test and any consequences of taking the test (i.e., how test scores will be used).
 - 2. Students should be informed about the test's different modalities (print, web, verbal) and available accommodations.

3. A process should be implemented to document that students have read, understood, and accepted the guidelines.

C. Administering and Scoring Tests

1. Detailed test administration procedures should be clearly outlined ahead of time and adhered to (time frame, use of books/notes).
2. Scoring procedures for evaluative measures (clinical performance, simulation, case analysis, etc.) should be delineated.
3. Interrater reliability should be regularly assessed.
4. Psychometric analysis should be used when possible to assure that the test is valid and internally consistent.
5. Methods of protecting the integrity of test items for standardized tests or other forms of testing, in which the items will be used in more than one context, should be clearly defined.

D. Reporting/Interpreting Test Results

1. Detailed norming information on standardized tests should be provided.
2. On tests used for predictive purposes, periodic evaluation of predictive validity should be included.
3. More than one mode of learning assessment should be used to make high-stakes decisions.

III. Recommendations to Achieve a Fair Testing Environment

- A. Convene a culturally and demographically representative group of faculty, students, and administrators to review your program's current high-stakes testing plans and policies.

Through the input of a diverse group of people affected by testing policies, new understanding about high-stakes tests and their consequences can be mutually discovered. All members of your review group should feel free to share their knowledge about the tests, their perceptions of how tests are used and their intended purposes, and the consequences of any change to testing policy.

- B. As a faculty, undertake a thoughtful and comprehensive review of the factors leading to the development and implementation of high-stakes testing in your program.

Program quality encompasses more than what is measured by licensure exam pass rates. The nursing education literature and stories from students and faculty alike reveal that high-stakes tests are often quickly

implemented in response to both internal and external pressures. The feeling of *having done something* can unintentionally divert faculty attention from other systems-related issues that bear on NCLEX-RN® pass rates and other measures of program quality. Factors such as admissions policies, instructional effectiveness, remediation and study requirements, and course-level assessments are all valid aspects of the educational process for review and improvement.

- C. Invite faculty or other experts with experience using high-stakes tests to provide feedback on how high-stakes tests are best used within the context of national guidelines, ethical considerations, and regulatory requirements.

Faculty members from other disciplines such as psychology, educational assessment, and psychometrics may have a longer history of using high-stakes tests in their educational practice. This could also be an opportunity to seek legal review of testing policies. This step is often overlooked during policy development but is increasingly important as schools seek to avoid costly and time-consuming legal battles, and the negative publicity that ensues.

- D. Until more formal studies are done, seek out and learn the practices of schools that have not needed to implement high-stakes testing, or that use tests in a non-high-stakes way, but still achieve excellent NCLEX-RN pass rates.

Across the nation—likely in every state—there are nursing education programs that maintain high NCLEX-RN pass rates. These programs admit very diverse students from a range of educational backgrounds, provide outstanding educational experiences, and have high retention rates. Students graduate from these programs, successfully pass the licensing exam, and enter the nursing workforce well prepared. Much could be learned about the effective practices and characteristics of these schools; their strengths are worthy of study and possible replication.

- E. Develop a communications plan for students and faculty that conveys essential information about your testing policy and practices.

Reassure students and faculty that local testing policies are aligned with NLN Fair Testing Guidelines, that there is strong psychometric support for using tests in fair and effective ways, and that testing policy, like other components of the overall assessment plan, considers the input of a variety of constituents, including students, faculty, and program leaders.

APPENDIX E

STANDARDS FOR TEACHER COMPETENCE IN EDUCATIONAL ASSESSMENT OF STUDENTS

Developed by the American Federation of Teachers
National Council on Measurement in Education
National Education Association

1. Teachers should be skilled in choosing assessment methods appropriate for instructional decisions.
2. Teachers should be skilled in developing assessment methods appropriate for instructional decisions.
3. Teachers should be skilled in administering, scoring, and interpreting the results of both externally produced and teacher-produced assessment methods.
4. Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and promoting school improvement.
5. Teachers should be skilled in developing valid student grading procedures that use student assessments.
6. Teachers should be skilled in communicating assessment results to students, parents, other lay audiences, and other educators.
7. Teachers should be skilled in recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information.

From Standards for Teacher Competence in Educational Assessment of Students. Developed by American Federation of Teachers, National Council on Measurement in Education, and National Education Association, 1990. This is not copyrighted material. Reproduction and dissemination are encouraged. Retrieved from <http://buros.org/standards-teacher-competence-educational-assessment-students>

INDEX

- abbreviations, test writing guidelines, 59
- accountability, 3, 351, 373
- accreditation models, 354–356, 373
 - evaluation of online programs using, 354
- achievement testing, 26, 30
- ACT scores, 154
- ADA. *See* Americans with Disabilities Act
- administering tests, 186–191, 394
 - in online courses, 200–203
- administration, evaluation of, 370
- affective domain
 - objectives taxonomy, 17–18
 - writing objectives, 16–18
- “all of the above” answers, 95
- alternate-forms reliability, 35, 43
- alternatives
 - in multiple-choice items, 80, 87–91, 381–382
- Americans with Disabilities Act (ADA), 324, 325
- analysis
 - in Bloom’s cognitive taxonomy, 15
 - level items, 149–150
 - See also* analyzing
- analytic scoring, 114–115
- analytic skills
 - assessment of, 152
 - testing of, 79
- analytical thinking, assessment of, 79
- analyzing
 - in revised cognitive taxonomy, 16
 - level items, 149–150
- anecdotal notes, 270
- anonymous assessment
 - grading system, 315
 - scoring of essay items and written assignments, 106–107, 116, 171
- answer(s)
 - changing, 64
 - key, 184–185
 - patterns, 182–183
- answer sheet, 75, 102, 180, 184, 185
 - machine-scored, 49
 - scannable, 53
- application
 - in Bloom’s cognitive taxonomy, 15
 - level items, 149
 - See also* applying
- application skills
 - assessment of, 105
 - testing of, 98, 149
- applying
 - in revised cognitive taxonomy, 16
 - level items, 149
- articulation, psychomotor skills, 19
- assessment, 3–5
- assessment-criterion relationship
 - considerations. *See* assessment validity
- assessment validity, 24
 - assessment-criterion relationship considerations, 29
 - consideration of consequences, 30
 - construct considerations, 26
 - content considerations, 25
 - defined, 24

- assessment validity (*cont.*)
 - historical perspectives, 24
 - influences on, 30
 - reliability, relationship with, 33
 - test blueprint as documentation of, 26
- assignment(s), 3–5. *See also* specific types of
 - assessment
 - assessment of, 165–172
 - bias, 312–315
 - legal aspects of, 323–325
 - online course, 200–203, 203–205
 - out-of-class, 125
 - results, reliability of, 33
- at-risk students, 348
- attitudes, student acceptance of, 17
- audio clips, in high-level learning
 - evaluation, 136
- autonomic reactivity, 65
- baccalaureate degree programs,
 - accreditation of, 352
- bar graph, 224
- behavioral techniques, test anxiety
 - reduction, 66
- belief system, evaluation and, 17, 257
- benchmarks, 368–369
- best-answer items, 87, 91–93, 98
 - guidelines for writing, 382
- best-work portfolios, 284
- bias
 - assessment, 312–315
 - sources of, 348
 - test, 325
- bimodal distribution, 225
- Bloom's taxonomy of cognitive domain, 15, 148
- blueprints, in test construction process, 48, 54–57
- browser security programs, 203
- calculations, short-answer items, 101
- carryover effect, 106–107, 171
- case analysis, 134, 269
- case method, 133, 282
 - examples, 132, 133, 134
- case presentations, 273
- case scenarios, 282
- case study, 133, 163, 282
- CAT. *See* computerized adaptive testing
- C-CEI®. *See* Creighton Competency Evaluation Instrument
- central tendency
 - error, 276
 - measures of, 228–229, 249
- certification examinations, 29, 52
- chart/exhibit item, NCLEX®, 128
- cheat sheets, 63
- cheating. *See also* online testing
 - low-technology forms of, 188
 - prevention strategies, 62, 186, 188–190, 201, 217
 - sanction for, 191
 - score reliability and, 39
- checklists, 271–272
 - description of, 271
 - design of, 272
 - performance evaluation, 271
 - sample, 272
 - uses of, 271
- “choice” items, in test construction, 52, 53
- CIPP model. *See* Context, Input, Process, Product model
- clarifying questions, 135
- clarity
 - in writing test items, 58
 - in written assignments, 162
- classroom evaluation, 9, 10
- clerical errors, in test construction, 64
- client needs framework, NCLEX®
 - care environment, safe and effective, 143
 - health promotion and maintenance, 143
 - physiological integrity, 144–157
 - psychosocial integrity, 143–144
- clinical competencies, 263–264. *See also*
 - clinical evaluation methods, rating scales

- clinical conferences, 286–288
- clinical evaluation, 257
 - concept of, 256–258
 - fairness in, 258–260
 - feedback, 261–263
 - formative, 258
 - versus grading, 257
 - media clips, 288–289
 - simulations, 278
 - subjective process, 256
 - summative, 258
 - tools. *See* rating scales
 - written assignments, 279–283
- clinical evaluation methods, 205, 210.
 - See also* rating scales
 - cases, 282
 - clinical conferences, 286–288
 - distance education, 279
 - form for evaluation, 290
 - group projects, 289–290
 - media clips, 288–289
 - observation, 269–278
 - for online nursing courses, 210
 - peer evaluation, 290, 291–292
 - portfolio, 284
 - rating scales, 273–276
 - selection factors, 267–269
 - self-assessment, 290–292
 - simulations, 269, 270, 273, 297
 - standardized patients, 268, 303
 - time factor, 270
 - written assignments, 279–283
- clinical evaluation tool. *See also* rating scales
 - guidelines for, 278
 - with multiple levels for rating
 - performance, 273
 - with two levels for rating performance, 275
- clinical judgment
 - evaluation of, 282
 - process, 122
- clinical learning, written assignments for, 163
- clinical observations, 164
- clinical outcomes, 263–264
- clinical performance, 205–210
 - faculty observation and evaluation, 208
 - preceptors, use of, 207–208
 - virtual site visits, 209
- Clinical Performance Evaluation Tool (CPET), 276
- clinical practice
 - competency of teachers, 361
 - evaluation of, 6
 - evaluation of students, 264
 - framework for test questions, 150
 - measurement of, 6
 - outcomes in, 253–255
 - student stress in, 260
- clinical practice grading systems
 - honors–pass–fail, 343
 - letter grades, 343
 - pass–fail, 345–349
 - satisfactory–unsatisfactory, 342
- clinical problem-solving assessment, 27
- clinical scenarios, 123, 127
- clinical setting, critical thinking in, 121
- clinical stations, Objective Structured
 - Clinical Examinations (OSCEs), 303
- clinical teachers, questions for
 - evaluating, 364
- Code of Fair Testing Practices in Education*, 321, 391–397
- coefficient alpha reliability estimate, 36
- cognitive component of test anxiety, 65
- cognitive domain
 - Bloom’s taxonomy of, 15, 148
 - sample verbs, 13, 16
 - taxonomy, 14–19
- cognitive learning, 15
- cognitive skills evaluation
 - case method, 132, 282
 - case study, 133, 282
 - distance education courses, 280
 - multimedia, 123
 - unfolding cases, 132, 282
- collaborative testing, 192–194
- Commission on Collegiate Nursing Education (CCNE) accreditation
 - process for online programs, 352

- communication skills
 - debates, 136
 - development of, 254, 402
 - journals, 280
 - writing assignments, 161
- competence/competency
 - demonstration of, 10
 - evaluation of, 10
 - for nurses in practice, 264
- completion items
 - characteristics of, 98
 - directions for, 179
 - test construction, 60
- comprehension
 - in Bloom's cognitive taxonomy, 15
 - level items, 148
 - See also* understanding
- computer software programs
 - grading, 349
 - item analysis, 183
 - learning management, 349
- computerized adaptive testing (CAT), 147
- computerized tests, 96
- concept analysis, 159, 163
- concept maps, 163, 280–282
- conciseness, in writing test items, 58
- concurrent validity evidence, 29, 42
- conferences
 - clinical evaluation, 286–287
 - criteria for evaluating, 288
 - distance education courses, 279
 - evaluation method, 286–287
 - learning plan or contract
 - development, 347
 - online, 282
 - postclinical, 124, 165, 286, 288
- confidentiality, 17, 319
- consequences of assessment
 - as validity evidence, 30
- construct validity evidence, 26–29
- constructed-response items, 53, 69
 - completion (fill-in-the-blank), 103
 - defined, 53
- content, in writing assignment, 166
- content validity evidence, 25–26
- Context, Input, Process, Product (CIPP)
 - model, 356
- context-dependent item sets, 122–123, 178
 - advantage of, 123
 - examples, 128–131
 - interpretive items on NCLEX®, 123–124
 - layout, 124
 - purpose of, 125
 - writing guidelines, 124–131
- context-dependent items, 61
- course assignments, 203–205
- course evaluation, 359–360
- cover page, 180
- CPET. *See* Clinical Performance Evaluation Tool
- creating
 - in revised cognitive taxonomy, 15–16
- credit-no credit grades, 333
- Creighton Competency Evaluation Instrument (C-CEI®), 300
- criterion-referenced clinical evaluation, 257–258
- criterion-referenced grading, 335–340
 - composite score computation, 337–338
 - fixed-percent method, 337–338
 - total-points method, 338–340
- criterion-referenced score interpretation, 7–8, 51, 230
- criterion-related validity evidence, 24, 29
- critical thinking skills
 - defined, 121
 - distance education courses, 280
 - evaluation of. *See* context-dependent item sets
 - writing objectives, 11–14
- crowding, avoidance of, 181
- C-SEI holistic tool, 301
- cultural bias, 314
- cultural competence, 254
- curriculum, 372
 - evaluation of, 359
- curve, grading on, 340–341

- data analysis skills, assessment of, 79
- data collection
 - in assessment process, 20
 - in evaluation process, generally, 8
 - in formative evaluation process, 9
- debates, 136
- debriefing, 287
- decision-making skills
 - development of, 17
 - evaluation of, 278
- decision-oriented program assessment
 - models, 356, 373
- delegation skills, 255
- deliberate practice, 18
- developing appropriate tests, 393
- diagnostic assessment, 3, 5
- dictation, 185
- DIF. *See* differential item functioning
- differential item functioning (DIF), 28, 313
- differential validity, 28, 313
- difficulty index, test item analysis, 234–235
- difficulty level of tests, 50–51
- directions
 - for writing assignments, 169
 - general, for tests, 179–180
 - item-format specific, 179
- disabilities, students with, 323
- discrimination
 - function of tests, 50
 - index, test item analysis, 234–235
 - in selection decisions, 312
- discussion, teaching format, 12, 134–136
- distractor analysis, test item analysis, 235–236
- distractors, multiple-choice tests, 80, 93–99, 98
- documentation
 - course failure, 348
 - observations of clinical performance, 270
- drafts, written assignments, 161–162, 165, 169
- effective evaluation, of program outcomes, 372
- effective teaching
 - administrator evaluation, 362, 373
 - clinical practice competency, 361, 373
 - evaluation of, 360–362
 - knowledge of subject matter, 361, 373
 - peer review, 215, 364–366
 - relationship with learners, 362
 - student ratings, 362–364
 - teacher, personal characteristics of, 362, 373
 - teaching portfolio, 366–367
 - teaching skills, 361
- electronic journals, 279
- electronic portfolios, 283–286
- eliminating items vs. adding points, 246–247
- emotional component of text anxiety, 65–66
- end-of-instruction evaluation, 21
- environmental conditions for testing, 186–187
- e-portfolio, as assignment, 164
- equivalent-forms reliability, 35, 43
- error score, 32, 37
- errors, in test construction, 30, 63
- essay items, 104
 - analytic scoring, 114–118
 - carryover effects, 106–107
 - criteria for assessing, 115
 - directions for, 179
 - essay item development, 106
 - extended-response essay items, 109
 - guidelines for writing, 387
 - holistic scoring, 113–114
 - limitations of, 105–106
 - organizing and outlining responses, 64
 - rater drift, 107
 - restricted-response essay items, 108
 - sample stems, 111
 - scoring, 113, 116–117, 388
 - student choice of items, 108–113
 - time, 107–108
 - unreliability in scoring, 106
- ethical issues, 319–323
 - Code of Fair Testing Practices in Education*, 321
 - importance of, 319
 - privacy, 319

- ethical issues (*cont.*)
 professional boundaries, violations of, 320
 test results, 320
 testing standards, 321
- ethnic bias, 313
- evaluating
 in revised cognitive taxonomy, 15–16
- evaluation. *See also* evaluating
 areas for, 368
 in Bloom's cognitive taxonomy, 15
 of courses, 359
 defined, 8
 formative, 9, 21, 132, 137, 203, 213, 258, 267, 268, 273, 329, 337
 instruction and, 10–11
 methods, 17
 objectives for, 8
 skills and, 8, 9, 21
 summaries, 346
 summative, 9–10, 132, 137, 203, 258, 267, 268, 329, 337
 types of, 8–11
- examinations. *See also* tests, testing
 certification, 52
 different forms of, 202
 licensure, 52
 proctored, 203
- extended-response essay items, 109
- external evaluators, 357, 373
- face validity, 26
- factual questions, 135
- faculty, evaluation of, 370
- failing grades
 clinical course, failure of, 349
 clinical practice, 345–349
 communication of, 341
 documentation requirements, 348
 effect of, 345
 for unsafe clinical performance, 348
 issues with assigning, 315–317
 problem identification, 346–347
 support services, 347–348
- failure, prediction of, 30, 322
- fair testing environment, recommendations for, 401
- fairness, 258–260, 311
- Family Educational Rights and Privacy Act of 1974, 319
- fatigue, impact on test-taking skills, 64
- feedback
 clinical evaluation, 261–263, 268, 269, 299, 303
 during conferences, 289
 in core competencies of nurse educators, 360
 in distance education, 280
 failing grades and, 346
 in online courses, 204–205
 performance evaluation, 276
 specific, 261
 5-step process, 263
 teaching evaluation, 364
 on written assignments, 160–161, 168, 172, 282
- feedback loop, in evaluation process, 11
- FERPA. *See* Family Educational Rights and Privacy Act of 1974
- fill-in-the-blank items, 102
 guidelines for writing, 386
- final grade, computation of, 337
- fixed-percent grading method, criterion-referenced grading, 337–338
- flaws, in test construction, 247, 250
- font selection, 185
- formal papers, in nursing course, 160, 162
- formative evaluation, 329. *See also* feedback
 clinical, 258
 defined, 9
 discussion as, 134–136
 documenting, 271
 in grading, 336
 implications of, 9, 213, 258, 281, 365, 366
 purpose of, 9
 simulations, 287
 by standardized patients, 303

- Formula 20 (K-R20)/Formula 21 (K-R21),
 - computation of, 36
- frequency distribution
 - characteristics of, 225–226
 - graphical display of, 224
 - of raw scores, 222–223
- frequency polygon, 224, 225
- gender, in differential item functioning, 28
- grade point average (GPA), 331
 - calculating, 333
- grading/grades
 - administrative purposes, 330
 - assessment bias, 315
 - clinical practice, 343
 - compression of, 317
 - consistent, 330
 - criterion-referenced, 335–340, 350
 - criticisms of, 330–332
 - on curve, 340–341
 - defined, 329
 - distinguished from scoring, 204
 - failing clinical practice, 345–349
 - framework selection, 335
 - group projects, 289–290
 - guidance and counseling purposes of, 330
 - importance of, 331
 - inflation of, 315–317, 332
 - instructional purposes, 330
 - learning contract, 347
 - letter grades, assignment of, 334–335, 350
 - as motivator, 331
 - norm-referenced, 340–341, 350
 - pass-fail, 343–345
 - purposes of, 329–330, 349
 - self-evaluation and, 331
 - self-referenced, 341–342, 350
 - software programs for, 349
 - spreadsheet application for, 349
 - summative evaluation, 9
 - types of systems, 332–334, 342, 350
 - written assignments as component, 279–283
- grammatical clues, 103
- grammatical errors
 - multiple-choice tests, 85
 - in test construction, 27
- group projects, 289–290
- group writing exercises, 164–165
- group-comparison techniques, 28
- growth and learning-progress portfolios, 284
- guessing answers, 64, 70, 233
 - correction for, 195–196
- half-length reliability estimate, 36
 - correction for, 36
- halo effect, 107, 276–277
- hand-scored tests, 53, 67, 184
- health care professionals, core
 - competencies, 254
- higher level learning, 119–120
- higher level thinking
 - assessment methods for, 132–137
 - context-dependent item sets, 122–123
 - problem solving skills, 120–121
- high-stakes assessments, 321–323, 399–402
- high-technology solutions to prevent
 - cheating, 200, 203
- histograms, test score distributions, 224, 225
- holistic scoring, 113–114
 - rubric, 113
- homogeneous content, matching exercises, 75
- honors–pass–fail grades, 333, 343
- hot-spot items, NCLEX® examination, 128, 146
- imitation level of psychomotor skill, 18, 19
- iNACOL. *See* International Association for K-12 Online Learning
- in-class writing activities, 164–165
- individual score, interpreting, 229–231
 - criterion-referenced interpretation, 230, 249
 - norm-referenced interpretations, 230, 249
 - results of standardized tests, 230–231
 - teacher-made tests, 230–231

- informal language, in test construction, 59
- information needs of test-takers, 395–397
 - student preparation for test, 62–63
- instructional design in online programs,
 - assessment of, 213, 354
- instructional process, evaluation skills
 - and, 9
- interactions, analysis of, 163
- internal consistency reliability evidence,
 - 36–37
- internal program evaluators, 357, 373
- International Association for K-12 Online Learning (iNACOL), 211
 - national standards for quality online courses, 211
- interpretive items, 61
- interrater reliability, 35, 37
- irrelevant data, in test construction, 61, 82–83
- item analysis, 231–243
 - computer software for, 183, 231
 - computer-generated report, 232
 - difficulty index, 232–234
 - discrimination index, 234–235
 - distractor analysis, 235–236
 - examining, data in context, 236–243
 - point biserial correlation coefficient, 235
 - item arrangement on tests, 177–178
- item bias, 313
- item formats in test construction
 - constructed-response items, 53
 - objectively scored items, 52–53
 - selected-response items, 53, 79
 - selection criteria, 51–52
 - subjectively scored items, 52–53
- item sequence, in test construction, 177–178
- jargon, avoiding use, 59
- Joint Committee on Testing Practices, 391–392
- journals, 159, 161, 164, 279–280
- judgment
 - about observations, 270, 276
 - clinical evaluation and, 253, 256, 261
 - multiple-choice tests, 79
 - pass–fail grading, 343
 - in test construction process, 26, 52, 54
- knowledge
 - acquisition, 15, 17
 - development, 15
 - in Bloom’s cognitive taxonomy, 15
 - level items, 148
- knowledge assessment, multiple-choice items, 80
- known-groups technique, 28
- Kuder–Richardson formulae, 36
- kurtosis, test score distributions, 225
- language, in test item writing, 58
- learners, positive relationships with, 362
- learning
 - assessment of in online programs, 201
 - disabilities, 184, 185, 192, 313, 314, 324
 - environment, significance of, 259–260, 290
 - management systems, 213
- learning contract, 347
- learning management systems, grading
 - systems in, 349
- learning outcomes
 - assessment of, 11
 - in teaching students, 21
- legal issues, 323–325
- legibility, significance of, 185
- length of test, 49–50, 52
- leniency error, 276
- leptokurtic distribution of scores, 225
- letter grades, assignment of
 - considerations in, 334–335
 - what to include in, 334–335
- licensure examination, 29, 30, 142
- linguistic bias, 314

linguistic modification for non-native speakers, 58

literature reviews, 161

logical error, 277

low-technology solutions to prevent cheating, 200–203

machine-scored tests, 49

manipulation level of psychomotor skill, 19

matching exercises

advantage of, 74

classification of, 53

components of, 69

directions, 179

disadvantages of, 74

examples, 75

guidelines for writing, 75–76, 378–380

measurement

criterion-referenced, 8

defined, 6

interpretation, types of, 7

validity, 24, 59, 323

measurement error, 32, 37, 184, 247

test security, 186

media clips, 136, 288–289. *See also* video clips/videotapes

median (Mdn), 227

median, score interpretation, 41

memorization, 48, 80, 105

memory aids, 63

mesokurtic distribution of scores, 225

modality, 225

mode (Mo), 227

motor skills, development of, 18

multimedia

clinical evaluation methods, 288–289

context-dependent items, 124

on NCLEX®, 123

multimodal distribution of scores, 225

multiple true–false items, 73

multiple-choice items, 80

advantages of, 79–80

alternatives, 87–91, 381–382

best-answer items, 87, 91–93, 98, 382

construction of, 51, 80–96

correct answer, 92, 382

design factors, 178, 186

directions for, 179

disadvantages of, 80

distractors, 80, 87, 93–99, 382

examples, 383–384

format, 86, 98

knowledge level, 148

negatively stated stems, 85

options arrangement, 183

parts of, 80, 81

purpose of, 80

scoring procedures, 53

stem, 81–87, 381

variation of items, 96

wording, 84, 85

writing guidelines, 81, 381–382

multiple-response items, 96–97, 385

alternatives, 97

computerized, 96

defined, 96

examples, 97, 385

order of responses, 97

writing guidelines, 97, 385

National Council Licensure Examination (NCLEX®)

ADA compliance, 325

administration of, 147

characteristics of, 52, 96

format, 102, 147

grade inflation and, 316

item preparation, varied cognitive levels, 147–150

predictive validity, 29, 42

predictors of success on, 29, 30, 154

preparing students for, 154–155

test plan, 142, 145

client-needs framework, 143

clinical practice framework, 150–154

cognitive levels, 145, 147

- National Council Licensure Examination (NCLEX®) (*cont.*)
 - integrated processes, 144
 - nursing process framework, 150
 - percentage of items
 - PN test plan, 145
 - RN test plan, 142
 - types of items, 146
- National Council of State Boards of Nursing (NCSBN), 142, 298
- National Council on Measurement in Education (NCME), 396
- National League for Nursing (NLN) *Fair Testing Guidelines for Nursing Education*, 399
- National Organization of Nurse Practitioner Faculties (NONPF), 208
- naturalization level of psychomotor skill, 19
- NCLEX®. *See* National Council Licensure Examination
- NCLEX test plans, 142
- NCME. *See* National Council on Measurement in Education
- NCSBN. *See* National Council of State Boards of Nursing
- needs assessment of learner, testing for, 20
- negative feedback, 259
- negatively stated stems, 85
- NLN. *See* National League for Nursing
- No Child Left Behind Act, 311
- “none of the above” answers, 95
- NONPF. *See* National Organization of Nurse Practitioner Faculties
- norm tables, 231
- normal distribution of scores, 226
 - for “grading on the curve”, 226, 340–341
- norm-referenced clinical evaluation, 257–258
- norm-referenced grading, 337
 - defined, 340
 - grading on the curve, 340–341
 - standard deviation method, 341
- norm-referenced score interpretations, 7, 229–230
- notes about performance, 270–271
- nursing care plan, 162, 163, 280
- Objective Structured Clinical Examination (OSCE), 271, 303–305
 - best-practice guidelines for, 305
- objectives
 - achievement of, 9
 - for assessment and testing, 11–14, 19–20
 - development of, 16
 - performance, 19
 - taxonomies of, 14–19
 - writing guidelines, 12
- observation. *See also* rating scales
 - in clinical evaluation, 269–270
 - significance of, 275
- on-campus/regional evaluation sites, 208
- online conferences, criteria for evaluating, 287
- online courses
 - assessment of, 211–213
 - assessment of teaching in, 214–215
 - assignments, assessment of, 200, 201, 203–205, 216, 363
 - clinical evaluation in, 205–210
 - evaluation of, 359
 - feedback in, 204–205
 - instructional design of, 213
- online education, defined, 199
- online education programs, assessing quality of, 216
- online instruction, program evaluation, 216, 365
- online learning, 137, 200
- online nursing program
 - assessment of, 355
 - critical elements for assessment, 354
- online teaching, assessment of, 214–215
- online testing, 200–203, 217
 - cheating prevention, 191, 200, 201, 217
 - conditions for, 187

- open-book tests, developing and
 - administering, 202–203
- open-ended questions, 135, 282
- open-ended response, 105, 113
- options
 - multiple-choice tests, 90, 93, 95
 - multiple-response tests, 97
- oral presentations, case analysis, 134
- organization
 - in affective domain, 18
 - in writing assignment, 166
- OSCE. *See* Objective Structured Clinical Examination
- outcome(s)
 - assessment, 3–5
 - clinical evaluation, 253
 - of clinical practice, 253–255, 263–264
 - criterion-referenced clinical evaluation, 8
 - evaluation, generally, 8–9
 - taxonomies, 14–19
 - use in assessment, 19–20
 - writing, 11–14
- papers, 282–283. *See also* written assignments
- “pass the writing” assignments, 165
- pass–fail grades, 333, 345–349
- peer evaluation, 215, 290, 291–292, 365
- peer review, 205, 282, 314, 364–366
 - faculty development for, 366
 - of online teaching, 215
- percentage-correct score, 230, 249
- percentile rank, 230
- performance problems, 346–347
- performance quality, 20
- personal bias, 277
- philosophical approaches, 373
- placement tests, 6
- planning for evaluation method, 267
- platykurtic distribution of scores, 225
- point biserial correlation coefficient, 235
- “pop” tests, 62
- population to be tested, 49
- portfolio
 - clinical evaluation, 283–286
 - contents of, 284
 - defined, 284
 - electronic, 283–286
 - evaluation of, 284
 - purpose of, 284
 - teaching, 366–367
 - types of, 284
- positive reinforcement, 262
- posttest discussions
 - conducting, 243–246
 - eliminating items vs. adding points, 246–247
- power test, 50
- preceptors, distance education, 287
- precision level of psychomotor skill, 11
- preclinical conference, 281, 286
- predictive validity, 29, 42
- premises, matching exercises, 75
- preparing students for tests, 61–66, 394, 395–396
- presentation skills, 136, 286
- printing guidelines for tests, 185
- privacy issues, 319
- problem-solving skills
 - assessment of, 19, 120, 148
 - context-dependent items, 144
 - ill-structured problems, 120–121, 142
 - improvement of, 159
 - well-structured problems, 142
- process, in assessing writing assignment, 166
- proctor, functions of, 188
- proctored examinations in distance education, 191, 203
- professional boundaries, violations of, 320
- program admission
 - candidate selection, 312
 - examinations, 6
- program assessment. *See also* program evaluation
 - curriculum evaluation, 357–360
 - ethics, 320

- program assessment. *See also* program evaluation (*cont.*)
 - online programs, 354
 - stakeholders, 30, 42, 357
 - standardized tests, use in, 26, 29, 368
 - teaching competencies, 361
 - teaching effectiveness, 362–367
- program development, evaluation and, 8
- program evaluation models, 356–357
 - accreditation, 354–356
- program outcomes, effective evaluation of, 372
- projects, evaluation of, 10
- proofread tests, 184
- psychometric issues, 323–324
- psychomotor domain
 - development of, 18–19
 - objectives taxonomy, 18
 - writing objectives, 18–19
- psychomotor skills
 - clinical evaluation feedback, 262
 - development of, 254
 - distance education courses, 268
- purchased tests, 41
- purpose of test, 49, 66

- QPA. *See* quality point average
- QSEN. *See* Quality and Safety Education for Nurses
- quality
 - of education, 3, 8, 21
 - improvement, as outcome of clinical practice, 253, 255
 - of teaching, 373
- Quality and Safety Education for Nurses (QSEN), 276
- quality point average (QPA), 333
- questioning, discussions, 134–136
- questionnaires, 320
- questions, during test, 179, 187–188
- quizzes, 69, 194, 329, 335, 337

- rater drift, 107, 277
- rating forms, 264, 276, 278, 346. *See also* clinical evaluation methods; rating scales
- rating scales, 273–276, 303. *See also* rating forms
 - applications, 276
 - benefits of, 273, 346
 - and clinical evaluation, 273–276
 - common errors of, 276
 - defined, 273
 - in evaluation process, 10
 - examples, 274
 - for final evaluation. *See* summative evaluation
 - guidelines for using, 278
 - issues with, 276–277
 - types of, 273–276
- raw scores, 194, 197, 222, 223
 - frequency distribution of, 223
- reading ability, 59, 64
- reading papers, 171
- recall
 - essay items and, 104
 - short answer, 101
 - test construction and, 141
 - testing of, 102
 - true–false, 70–74
- receiving, in affective domain, 17
- recording observations, methods for, 271. *See also* checklists
- reflective journaling, 168
- relaxation techniques, for test anxiety, 66
- reliability
 - alternate-forms, 35, 43
 - assessment reliability, 31, 33
 - consistency of ratings, measure of, 37
 - decay, 277
 - defined, 31, 42
 - equivalence, measure of, 35–36
 - equivalent-forms reliability, 35, 43
 - error, relationship to, 32
 - estimating, 33, 35

- grading system and, 329
- influences on, 31, 38
- internal consistency, measure of, 36–37, 43
- scorer, 37
- significance of, 31
- stability, 35
- test-retest, 35
- validity, relationship with, 33
- remedial learning, 5, 347
- remediation, 154, 346
- remembering
 - in revised cognitive taxonomy, 16
 - level items, 148
- reproducing tests, 185–186
- research papers, 162
- responding, in affective domain, 17
- Respondus™, 203
- restricted-response essay, test construction, 53, 108
- review sessions, 66
- rewrites, written assignments, 161–162
- rubric, 167
 - analytic scoring, 114
 - for assessing conferences, 289
 - for assessing group projects, 291–292
 - for assessing papers, 166, 172
 - for assessing portfolio, 289
 - benefits of, 165
 - holistic scoring, 113–114
 - sample scoring rubric, term paper, 167
 - scoring, 205
 - written assignments, 166, 172
- “rule of C”, 316
- SAT scores, 154
- satisfactory–unsatisfactory grades, 333, 342
- score distribution
 - test. *See* test score distributions
- score interpretation
 - criterion-referenced, 7–8, 20, 230, 233, 249
 - norm-referenced, 7, 230–231, 233, 249
- scoring, 53–54, 194–196
 - analytic, 114
 - components of, 6, 113
 - computerized, 96
 - correction for guessing, 195–196
 - defined, 205
 - essay tests, 105
 - facilitation of, 178, 181, 182
 - inconsistency of scores, 31
 - inflation of, 315–317
 - influential factors, 29
 - measurement validity, 24
 - multiple-choice tests, 27, 40
 - objectively scored test items, 52–53
 - reading papers, 168
 - rubric, 205
 - subjectively scored items, 52–53
 - suggestions for, 116–117
 - unreliability in, 106
 - weighting items, 195, 335
- scoring tests, 194–196, 392, 394
- security issues
 - cheating prevention, 188–191
 - online testing, 202
 - in test reproduction, 186
- selected-response test items
 - characteristics of, 53
 - effectiveness of, 69
- selecting appropriate tests, 393
- self-assessment, 255, 290–292
- self-esteem, tests and grades effects on, 317–318
- self-evaluation, 292, 331
- self-referenced grading, 335, 342–349, 350
- self-study, program, 352
- SEM. *See* standard error of measurement
- severity error, 276
- short cases, advantages of, 282
- short papers, 163, 165, 282
- short written assignments, 137
- short-answer essays, 108

- short-answer items, 82, 96. *See also*
 - fill-in-the-blank items
 - examples, 103–104, 386–387
 - guidelines for writing, 102–103, 386–387
- simulation-based assessment, guidelines
 - for, 302
- simulations
 - for assessment, 273, 297–302, 302
 - characteristics of, 254, 268
 - clinical evaluation usage, 273
 - Objective Structured Clinical Examination (OSCE), 271, 303–305
 - standardized patients, 271, 303, 306
 - types of, 273–276
- skills
 - development, 4
 - in teaching, 360, 363, 373
- slang, test writing guidelines, 59
- small-group writing activities, 164–165
- social issues, 311–312
 - assessment bias, 312–315
 - grade/test score inflation, 317–318
 - occupational roles, 312
 - self-esteem, influential factors, 317–318
 - testing as social control, 318–319
 - types of, 312
- social loafing, 193
- spacing, in test design, 196
- Spearman–Brown double length formula, 36
- Spearman–Brown prophecy formula, 36
- speeded test, 50
- spelling errors, in test construction, 27, 246
- split-half reliability, 36, 43
- stability, measure of, 35, 43
- standard deviation (SD), 228
 - norm-referenced grading, 341
- standard error of measurement (SEM), 37
- standardized patients, 214, 268, 271
- standardized test manuals, 231
- standardized tests
 - ACT, 154
 - characteristics of, 50, 88
 - equivalent-form estimates, 35
 - National League for Nursing, 399, 402
 - program assessment, 26, 357
 - results of, 230–231
 - SAT, 154
 - scores of, 29
 - Standards for Educational and Psychological Testing*, 25, 314
 - Standards for Teacher Competence in Educational Assessment of Students*, 403
- stems
 - essay items, 111
 - multiple-choice items, 81–87, 98, 381
- storage of test materials, 186
- stress
 - in clinical practice, 260
 - principles of, 261–263
 - reduction strategies, 318
- structural bias, 314
- student
 - characteristics, 31–35
 - with disabilities, assessment of, 324–325
 - placement process, 6
 - preparing for test, 61–66, 394, 395–396
 - ratings, 362–364
 - records, 319, 320
 - study skills, 317
 - teaching effectiveness, evaluation of, 362–367
 - test anxiety, 64–66, 155, 317–318
 - test-taking skills, 38, 63–64
 - test-taking strategies, 318
- student achievement assessment
 - methods, 358
- student characteristics, assessment results, 31–35
- student evaluation of teaching. *See* teacher, evaluation of
- student–faculty interaction, 354
- study skills, 317
- summative assessment, 298, 303
- summative evaluation
 - clinical, 258, 268, 273
 - defined, 9

- in grading, 329, 336, 343, 345
 - tests, 10
 - “supply” items, in test construction, 70, 77
 - support services, 347–348, 354
 - supportive environment, 259–260
 - syllabus, 337, 345
 - synthesis
 - in Bloom’s cognitive taxonomy, 15
 - See also* creating
 - systematic program evaluation (SPE)
 - components of, 368
 - criteria, content of, 370
 - evaluation areas, content of, 370–372
 - evaluation framework, 368
 - systematic reviews, 163
 - systems-oriented models, 368
- take-home tests, 202
- taxonomies, 14–19
- teacher
 - evaluation of, 370–372
 - personal characteristics of, 362, 373
 - relationship with students, 268
- teacher-constructed test, 202
 - blueprint for, 67
 - interpreting results of, 229–230
 - length of, 49–50
 - preparing students for, 40, 61–66
 - time constraints, 52
 - validation for, 27
- teacher–student relationship, 259, 268
- teaching
 - assessment of, 363
 - online, 200–203
 - peer review of, 215
 - plan, 163
 - portfolio, 366–367
 - skills in, 361
 - student evaluation of, in online learning, 215
- teaching to the test, 317
- teaching–learning process, 10, 362, 366
 - evaluation of, 359–360
- technical skills, development of, 268
- technological skills, development of, 18
- term papers, 162, 167, 283
- test administration
 - answering questions during test, 187–188
 - cheating prevention, 188–191
 - collaborative testing, 192–194
 - collecting test materials, 192
 - conditions, 39
 - cost of, 40
 - directions, 40, 188
 - distributing test material, 187
 - environmental conditions, 186–187
 - ethical issues, 321
 - and scoring, 401
 - time factor, 50
- test anxiety, 64–66, 155, 178, 318
- test blueprint, 54–57
 - body of, 54
 - column headings, 54
 - content areas, 54, 67
 - defined, 54–57
 - elements of, 54
 - example of, 54
 - functions of, 56
 - review of, 55
 - row headings, 54
 - for students, 67
- test construction
 - checklist for, 62
 - content areas, 54
 - cost of, 40
 - difficulty level, 50–51
 - discrimination level, 50–51
 - item formats, 49, 51–53, 155
 - population factor, 49, 66
 - scoring procedures, 53–54
 - test blueprint, 54–57, 67, 145
 - test items, development of, 12
 - test length, 49–50, 52, 66
 - writing test items, guidelines for, 57–61, 377–388
- test contents, 26, 55, 67

- test design rules
 - answer key, 184–185
 - answer patterns, 182–183
 - cover page, 180
 - crowding, avoidance of, 181
 - directions, writing guidelines, 179, 188
 - item arrangement in logical sequence, 177–178
 - number of items, 183–184
 - options arrangement, 184
 - proofreading, 184
 - related material, arrangement of, 181–182
 - scoring facilitation, 182
- test developers, 391, 392, 394, 395
- test development, and implementation, 400
- test length, 49–50, 52, 66
- test materials
 - collecting, 192
 - distribution of, 187
 - security of, 186
- test planning
 - item formats, 49
 - preparing students for tests, 61–66
 - purpose and population, 49
 - types of items, 123
- test reproduction
 - duplication, 185
 - legibility, ensure of, 185
 - printing, 185
- test results
 - communication of, 321
 - reporting and interpreting, 392, 395, 401
- test score adjustments
 - effects of, 246
- test score distributions, 221–226
 - characteristics of, 225
 - graphic depictions of, 224, 225
 - shape of, 226
- test scores, interpreting, 221–226, 249
 - frequency distribution, 222, 224
 - measures of central tendency, 226–228
 - measures of variability, 228–229
 - raw scores, 222, 223
- test users, 391, 392, 393–395
- testing
 - concept of validity in, 25
 - definition of, 5
 - objectives of, 11–14, 19–20
 - purpose of, 5–6
- testing resources, for nurse educators, 389
- test-item bank
 - developing, 247–248
 - published, 248
 - sample information, 248
- test–retest reliability, 35
- tests
 - administering and scoring, 394
 - developing and selecting appropriate, 393
- test-taking
 - skills, 38, 63–64
 - strategies, 318
- testwiseness, 63
- time constraints, test construction process, 52
- time frame for program evaluation, 369
- total-points method, grading system, 338–340
- true score, 32
- true–false items
 - construction of, 50, 71–72
 - described, 70
 - item examples, 73, 182, 378
 - limitations of, 70
 - multiple, 73
 - scoring procedures, 53–54, 182
 - variations of, 72–74
 - writing guidelines, 71–72, 377–378
- “Truth in Testing” laws, 319
- typographical errors, 177, 184, 185
- unannounced tests, 318
- understanding
 - in revised cognitive taxonomy, 16
 - level items, 148
- unfolding cases, 133, 282
- unimodal curve, 225
- unsafe clinical performance, 348
 - policy, 349
- usability, significance of, 39

- validity
 - concept of, 23, 24
 - grading system and, 329
 - importance of, 26
 - influences on, 30
 - legal issues, 324
 - test blueprint and, 56
 - test construction and, 60
- values/value system
 - clarification strategies, 268
 - determination of, 254
 - development, 17, 19
 - internalization of, 18
 - organization of, 18
- valuing, in affective domain, 18
- variability, measures of, 228
- video clips/videotapes, 12, 136–137, 288.
 - See also* media clips
- visual disabilities, 185
- visual imagery, in test anxiety, 66
- weighting
 - grading software programs, 335
 - letter grade assignment, 335
- weighting system, 195
- worry, in test anxiety, 65
 - “catastrophic fantasies”, 65
 - “competitive worry”, 65
- word choice, in test construction, 60
- writing activities. *See also* written assignments
 - in-class and small-group, 165
 - for postclinical conferences, 165
- writing in the discipline and writing-to-learn activities, 160–161
- writing skills
 - development of, 159
 - improvement strategies, 283
 - of student, 27
- writing structure checklist, 170
- writing style, 162, 165, 166
 - and format, 167
- writing test items
 - within framework of clinical practice, 150
 - varied cognitive levels, 147
- writing test items, guidelines for, 57–61, 377–388. *See also* test construction
 - chart/exhibit, 123
 - essay, 110–113, 387
 - general rules, 57–61
 - hot spot, 123
 - matching exercises, 74–76, 378–379
 - multiple-choice, 81, 381–382
 - multiple-response, 97, 385
 - short answer, 101, 386
 - true–false items, 71–72, 377–378
- writing-to-learn activities, writing in discipline and, 160–161
- written assignments
 - assessment, 166
 - case method, 282
 - case study, 282
 - characteristics of, generally, 12
 - clinical evaluation, 279
 - concept maps, 280–282
 - drafts, 161–162, 165, 169
 - evaluation/grading, 168
 - feedback, 160–161, 168, 172
 - formal papers, 160, 161, 169
 - journals, 159, 161, 164, 168, 279–280
 - nursing care plan, 280
 - papers, 282–283
 - peer review, 165
 - purposes of, 159–162
 - rewrites, 161–162
 - rubric. *See* rubric
 - short written assignments, 137
 - types of, 162–165, 172, 279
 - unfolding cases, 132–134, 282