Speaker's Notes for Lesson 3	
1	Welcome to Unit 3. This unit examines Assessment and Evaluation Strategies. The next three slides of the presentation include the test plan objectives for this content. There will not be any voiceover for these slides. The voiceover presentation begins on slide 5. For ease of listening, references are not cited in the voiceover but are available within the speaker's notes.
2	
3	
4	
5	Assessment and evaluation are critical to the educational process. Through assessment and evaluation we are able to determine student and curricular needs, effectiveness of teaching and learning activities, and demonstrate achievement of program outcomes. Before we delve too far into assessment and evaluation, it is important to define some terms so that we are all operating from the same point of reference. Billings & Halstead (2012) indicate that assessment refers to measures that provide information before, during, and after participation in a learning activity or program (as cited in Baumlein, 2015, p.50). Assessment is broad and can be formal, informal, or inferred. Formally one may assess through measurement. Informally, one may assess by asking a question in class about content or by having students complete a minute paper related to their class learning. We might infer assessment. Consider students who are entering a class that had pre-requisites. Knowing that certain classes were completed prior to the start of another class may give one some sense of what students coming into class "know" and what one will build upon. Oermann & Gaberson (2009) indicated that measurement is a process of assigning numbers to represent student performance (as cited in Baumlein, 2015, p. 51). These numerical values may be reported as absolute values (i.e. 90 points out of 100 or 90%) or as relative values (i.e. 75th percentile). Measurement can be reported as norm referenced or criterion referenced. Norm referenced means that we compare one student's performance to others in a group. Performance of all students reflects the bell curve and individual student performance is reported as a percentile. Norm referenced interpretation allows one to rank students in a group (may be their own group, a class, or a large population who may have taken the same exam). Norm referencing is helpful in terms of its predictive value (Bourke & Ihrke, 2016). For instance, one may find that students applying to a nursing program who score at the

	60th percentile or better are generally successful in nursing school. This value may be a requirement for program admission. A disadvantage of norm referenced measurements is that the score is not absolute and it does not provide a reliable means of judging student performance in relation to learning objective achievement (Baumlein, 2015). In norm referenced measurements, student scores will not demonstrate the absolute number of questions answered correctly. Criterion referenced scores are absolute measures of performance and are not influenced by how other students do on the same assessment. Student achievement is based on preset criteria (Baumlein, 2015). If there are 100 questions and the student answers 93 correctly, their score is 93%. Criterion referenced measurements compare student performance against the learning objectives and can be used to interpret meeting those objectives. Criterion referenced measures are most commonly used in higher education in the form of tests, writing assignments, clinical performance, and licensing/certification exams. Evaluation is more formal, allowing us to appraise quality (Baumlein, 2015). The purposes of evaluation may include facilitating learning, diagnosing a problem, promoting decision making, and judging effectiveness (Bourke & Ihrke, 2016). The timeframe of the evaluation process can be formative or summative. Formative evaluation occurs during the program of learning to determine if progress is being made toward the outcomes. Formative evaluation data allows us look at pieces of the learning process and revise if necessary. For instance in a formative clinical evaluation we are able to determine if a student is moving consistently to the clinical outcomes and if not, remediate before it is too late for the student to be successful. Formative evaluations are a means of monitoring student progress and improving student performance before the course has ended (Gronlund & Waugh, 2009, as cited in Bourke & Ihrke, 2016). Summative evaluation occurs at the end of instruction and determines if the student has met the intended outcomes. A final course grade would be considered a summative evaluation (Baumlein, 2015).
6	Consider how program standards influence admission, progression, and graduation policies. These policies are important to student success and should have a foundation in research and best practices, and support program goals.

	Starting with admission policies, one must consider attributes students should possess to be successful in the program of study. Care must be taken to ensure that admission policies are not discriminatory in nature. Common admission requirements might include a minimum grade point average (GPA) in previous education, completion of pre-requisite coursework; a minimum score in a college readiness exam like the ACT or SAT or a standardized entrance test like the Nurse Entrance Exam or Teas Nurse Entrance Exam. The intent of these requirements is to provide objective criteria upon which to base admission decisions. GPA requirements should be considered carefully as grade inflation in secondary and post-secondary schools may impact the objectiveness of this indicator. For this reason, college entrance exams may be a more valuable indicator of student future performance (Christensen, 2016). These requirements should be consistently applied and based on evidence that the criteria outlined is reflective of appropriate qualifications for potential students. Progression policies must be fair with clear rationale as to how these policies support program goals. Common elements of progression policies include minimum GPA in coursework, drop policies, conditions for dismissal or re-admittance to the program, etc. Progression policies must be known to students and readily accessible. For this reason, progression policies (or a link to these policies) are often found in course syllabi, program handbooks, and college catalogs. Application of progression policies should always be based on data (Baumlein, 2015) and evidence. Graduation policies outline the requirement for program completion. This might include completion of program coursework, meeting financial obligations to college, completion of program requirements such as a portfolio, etc. Some programs include high-stakes testing as part of the program requirements for progression and graduation. High stakes testing might include a minimum score on tests designed to demonstrate preparedness for the NCLEX. The National League of Nursing Board of Governors' white paper entitled *The Fair Testing Imperative in Nursing Education* addresses ethical and legal concerns to high stakes testing that impacts progression and graduation. The NLN indicates that high stakes testing should not be the sole factor for a student not progressing or graduating from a nursing program (as cited in Baumlein, 2015). Instead multiple factors and sources of data should be utilized to determine student progression and graduation. These may include but are not limited to course exams, assignments, clinical experience evaluations, and other learning activities.

	Program standards are not without legal and ethical implications. First and foremost, programs standards should be equitable and be consistently applied. Students must be aware of program standards and these should be readily accessible to the student. Students are afforded the opportunity to file a grievance if they feel a policy has been inconsistently applied and or file suit. Consistent application of program standards based on objective data is the best means to ensure the rigor of the program as well as protect the school, student, faculty, and the public.
7	The cornerstone of assessment and evaluation is the intended learning objectives and outcomes identified through the program of study, courses, and individual class sessions. Careful attention must be observed when selecting an assessment or evaluation method to ensure that the method selected measures the learning objective or outcome at the intended level of performance. Three domains of learning, each having 5-6 levels, serve as the basis for learning. Early coursework may focus on the lower ends of the domains, while coursework at the end of a program of study may be more focused on the higher levels of the domains. However, even at the end of the program of study, if content that is new to the learner is introduced, learning will start at the bottom of the domain's levels (Kirkpatrick & DeWitt, 2016). The cognitive domain focuses on levels of knowledge. The hierarchical levels build from simple to complex. These levels are remembering, understanding, applying, analyzing, evaluating, and creating (Kirkpatrick & Dewitt, 2016). As you can see to move through the levels of this domain, one must complete the prior level. For instance you must be able to remember before you can understand; or understand before you can apply. The psychomotor domain addresses manual skill development and is commonly used in practice or simulation learning objectives. There are five levels in the psychomotor domain that include, from simple to complex, imitation, manipulation, precision, articulation, and naturalization (Kirkpatrick & DeWitt, 2016). Consider this progression from simple to complex in the psychomotor domain in relation to swimming. Initially our efforts at swimming may result in swallowing quite a bit of water and someone cuing to "kick your feet" and synchronizing your arm movements. As we practice, we become more skillful in swimming where we are able to swim across the pool without these cues but yet require availability of guidance. This would be considered the precision level. Further practice allows us to be more skilled in swimming to the point where it is a natural and

	fluid movement where we give little thought to how our body is moving through the water. This would reflect the naturalization level. Relating this analogy to nursing practice, chances are we aim for our students to reach the precision level with less utilized skills. In the precision level they can perform the skill in simulation or practice without faculty cues. In those skills that are utilized often, students may be able to reach the articulation or naturalization level of the psychomotor domain. The affective domain addresses emotions, values, and feelings. This domain creates the greatest challenge in teaching and measuring. Think about the discussions you may have had about how to measure caring. We know it when we see it but it is difficult to quantify or measure. Levels in the affective domain, from simple to complex, include receiving, responding, valuing, conceptualizing and organizing, and internalizing the values concept (Kirkpatrick & DeWitt, 2012). At the lower level of the affective domain, the student is a passive recipient of information. Consider how faculty may stress the importance ("value") of sterile technique. As students learn more about patient care, the need for sterile technique in certain settings, and the possible consequences of failure to use sterile technique, their importance they assign to this concept begins to deepen and become internalized as value that they would follow in all circumstances. Once the learning objectives and outcomes have been developed, faculty will select means to measure achievement. Validity means that we actually measure what we set out to measure. To do so, three attributes should be present. These are relevance, accuracy, and utility. Relevance means that the assessment measures the educational objective or outcome as directly as possible (Kirkpatrick & DeWitt, 2016). Accuracy indicates that the learning objective is measured precisely (Kirkpatrick & DeWitt, 2016). For instance if the learning objective was to "identify foods high in iron" and our assessment was to plan a balanced meal, one can see that the assessment strategy would have low relevance and accuracy as it is not a direct or precise measure of being able to identify foods high in iron. Lastly, valid instruments have utility. This means that it may provide formative and summative results allowing for ongoing improvement and final evaluation measures (Kirkpatrick & DeWitt, 2016). Valid test questions might be used as part of a unit exam allowing faculty to determine areas where additional instruction is needed for students. That same question may be used within a final comprehensive exam to ensure that learning did occur.
8	Kirkpatrick & DeWitt (2012) outline six strategies in selecting an assessment or evaluation tool. Let's expand a bit on several of these

	strategies. The purpose of the assessment is largely driven by the learning objectives that indicate the type of behavior to be assessed (cognitive, psychomotor, or affective). Cognitive learning is typically assessed through writing. This may be papers, test completion, essays, etc. Assessment in the psychomotor domain generally involves simulation and clinical. Assessment in the affective domain is often found in the clinical setting where students have the opportunity to demonstrate the values of nursing through interactions and caring for patients (Kirkpatrick & DeWitt, 2016). Reliability of assessments is critical, particularly when grading has some subjectivity (i.e. clinical performance, papers, presentations, etc.). When considering how an assignment will be graded, one must consider intra-rater reliability and inter-rater reliability. Intra-rater reliability reflects an individual's ability to be consistent in how assignments are graded with all students. Inter-rater reliability reflects consistency of grading between two or more faculty members (Kirkpatrick & DeWitt, 2016). Use of clear grading criteria and rubrics can assist in establishing both forms of reliability. Rubrics, particularly those that delineate levels of performance, allow for more consistent grading. Inter-rater reliability should be established before assessment is graded. To establish inter-rater reliability, graders should independently rate performance based on the grading criteria. Once this has been completed, inter-rater reliability is calculated by taking the total number of agreements divided by the total number of agreements and total number of disagreements. Inter-rater reliability should be >70% although higher is better (Polit & Hungler, 1999, as cited in Kirkpatrick & DeWitt, 2016). After the completion of an assessment it is important to reflect on the effectiveness of the assessment. According to Kirkpatrick & DeWitt (2016, p. 401), some questions to consider are: 1. Was the strategy an effective use of resources? 2. Was there adequate data to determine of learning objective or outcome was met? 3. Were there any problems with implementation? If so, what revisions are recommended?
9	While detailed discussion on test construction is beyond the intent of this course, discussion of test blueprints and an overview of construction are worth some time. Test blueprints are maps that connect content to outcomes. Content that may be included on test blueprints include relationship of each

	question to the learning objective (domain and level in the objective); nursing process; NCLEX test plan category, and weighting based on time spent on this objective within the class itself (Billings, 2016). Test blueprints are designed to be internal documents that assist in establishing test validity (Baumlein, 2015). In relation to writing test items, Billings (2016) provide some considerations related to validity and reliability: 1. True false questions: reliability is low as one has a 50% chance of guessing the correct response; valid for lower levels of the cognitive domain 2. Matching questions; reliability may be low as these types of questions are difficult to write without giving clues; valid for lower levels of the cognitive domain 3. Short answer/fill-in the blank: minimizes guessing which increases reliability; can measure lower to middle levels of the cognitive domain 4. Multiple choice questions: reliability is increased with multiple answer multiple choice questions; when carefully constructed can measure higher levels of the cognitive domain One last note to consider is use of test banks. Test bank questions are written to reflect the learning objectives/outcomes of the book. If using test bank questions, it is important to determine the validity of these test questions based on your course's learning objectives and outcomes.
10	While there are a number of measures available in analyzing test results, three parameters will be examined here. Item difficulty level (p value) allows the educator to determine if a question was too easy, too difficult, or on target. Difficulty level is calculated for each question and reflects the percentage of students who answered the question correctly. This could be reported as a decimal or a percentage. A difficulty level between 0.3-0.9 is considered acceptable (Morrison, 2010 as cited in Baumlein, 2015). Item discrimination allows faculty to discriminate between those who knew the content and those who did not. Item discrimination compares each student's question performance with their overall test performance (Billings, 2016). The two measures of item discrimination are the item discrimination ratio (IDR) and point biserial correlation coefficient (PBCC). Let's start with the IDR. To do so, we examine the top 27% of test scorers and the lowere 27% of test scorer's performance on individual test questions. The IDR is calculated by taking the percent of the top 27% answered a question correctly minus the percent of the bottom 27% who answered the

	question correctly. An acceptable level for the difference is 25% or greater. The PBCC is considered the most accurate reflection of item discrimination (Baumlein, 2015). PBCC is calculated using test software. The PBCC score can range between 1.00 to -1.00. A positive number indicates higher scoring students answered an individual question correctly more often than lower-scoring students. PBCC indices greater than 0.3 are considered good and greater than 0.4 are considered very good. Any PBCC <0.2 indicates the question should be rewritten related to low discrimination. The PBCC score can be used to evaluate the quality of question distractors; a negative score is desired as this indicates the lower scoring students selected this distractor more commonly than higher scoring students (Billings, 2016). Lastly, Billings (2016, p.436) states the reliability "refers to the ability of a test to provide dependable and consistent scores". Reliability scores (Kurder-Richardson or KR-20) range from 0 to 1.00. A reliability co-efficient of 0.6 or higher is considered acceptable for teacher made exams (Baumlein, 2015) and 0.7 or higher for standardized exams. The higher the score the greater the reliability. Measures of reliability assume that all items on the test are about the same difficulty and there were not external factors (i.e. not enough time) that negatively influenced students' ability to complete the exam.
11	Please review the example of difficulty level (p value) and PBCC for the statistics outlined on this slide for two questions. As you can see the p value for both questions fall within the acceptable range of 0.3-0.9. the overall PBCC for both questions are also acceptable as both are greater than 0.3. Let's examine the distractors. Let's start with question #2. as you can see all the distractors in a negative value indicating that those who scored lower on the exam selected those responses more often than those who scored higher on the exam. But now let's look at distractors in question #1. Distractor A was not selected by any student as indicated by the 0.. This is not an effective distractor and students are eliminating this choice. This increases the chance of students guessing the correct response as now there are three choices to rather than four choices. Distractor A should be rewritten. Another concern is Distractor D. This positive score means that higher scoring students were more likely to select this wrong response than lower scoring students. This indicates poor discrimination and should be rewritten.

	Lastly, look at the reliability or KR-20 score. This score in question #1 is 0.56 which falls below the desired range of 0.6 or higher. Likely this is due to the poor discrimination in the distractors and the lower overall discrimination of the question. The KR-20 for question #2 is quite good as it is well above the minimum score of 0.6.
12	A key factor to any assessment or evaluation measure is adequate communication about the evaluation criteria. Students must be clear on policies that impact progression and graduation. In classroom assessments, students should have a clear understanding of what learning objectives are to be measured and how these will be measured. Providing students with rubrics or other types of evaluation criteria prior to the assessment allows students to better prepare. Feedback should be constructive, timely, and thoughtful. Beach & Marshall, 1991 (as cited in Kirkpatrick & DeWitt, 2016) outline seven components to responding to writing assignments (p.412). These include: 1. Praising—positive reinforcement increases chance for seeing this again 2. Describing—provide reader-based feedback that includes your own reaction to their line of thinking and explanation. 3. Diagnosing—determine the students level of knowledge, attitudes, abilities, and needs 4. Judging—evaluate the completeness, validity, and insightfulness of work 5. Predicting and reviewing growth—provide direction of how student may improve related to needed areas of growth 6. Record keeping—keep notes about how student performs across time 7. Recognizing and praising growth—provide positive feedback about improvement even if student still has more room to grow It should be clear to students how a writing assignment will be or was graded. This may be through comments within a paper or comments upon the rubric correlating their level of work with that of the criterion desired. Providing clear and specific examples of where the student excelled and where additional growth is needed is critical. Feedback must be timely. Students should not submit another assignment before the first assignment feedback has been provided allowing them to improve. This unit has addressed the evaluation process. Please review the reference list for further resources to enhance your understanding.

	After adequate review, please proceed to the post test. A minimum of 90% must be achieved on the post test to progress to the next unit.