JNPD Journal for Nurses in Professional Development • Volume 31, Number 1, 2–8 • Copyright © 2015 Wolters Kluwer Health, Inc. All rights reserved.

Refine Test Items for Accurate Measurement

CE
3.0 ANCC
Contact
Hours

Six Valuable Tips

Karen Siroky, MSN, RN-BC ❍ Bette Case Di Leonardi, PhD, RN-BC

Nursing Professional Development (NPD) specialists frequently design test items to assess competence, to measure learning outcomes, and to create active learning experiences. This article presents six valuable tips for improving test items and using test results to strengthen validity of measurement. NPD specialists can readily apply these tips and examples to measure knowledge with greater accuracy.

Nursing Professional Development (NPD) specialists continuously devise and improve upon approaches to assess and validate competency. Competency embraces the cognitive, affective, and psychomotor domains of learning and performance. Tests function as one indicator in competency management models by measuring the cognitive domain: a clinician's knowledge base or competence.

Although nursing examinations have begun to introduce alternatives to multiple-choice items, the multiple-choice item remains prevalent (Sutherland, Schwartz, & Dickison, 2012). Multiple-choice test items measure competence/knowledge and not competency/performance. However, test results make a more valid contribution to competency assessment/validation when test developers sharpen their focus on the knowledge pertinent for practice and frame test items in a practice context.

Despite a careful test planning process, flaws in test item construction can distract from accurate measurement and threaten validity (Haladyna, Downing, & Rodriguez, 2002; McDonald, 2013; Oermann & Gaberson, 2013). Flaws in test item construction draw the test taker's focus away from the point of the question by creating "noise." In this sense of the word, noise includes any features irrelevant to the intended measurement that make a question more difficult to decipher and answer correctly. NPD specialists need to eliminate noise to the greatest extent possible to assure that their test items precisely measure the knowledge/competence they intend to measure. NPD specialists frequently design test items to assess competence. In the authors' organization, the competency model includes knowledge/competence assessment examinations along with skills checklists, letters of reference, background checks, and ongoing performance appraisals to document competence and competency of nurses in a wide range of specialties and allied health personnel.

NPD specialists also develop test items to measure learning outcomes and to create active learning experiences. When designing posttests and interactive learning methods, learning objectives guide the selection and allocation of questions. Games in live sessions and interactive features in online courses use questions to engage the learner. Well-written questions give the learner practice in applying course content to realistic practice situations, that is, to put the objectives of the course into action.

When constructing tests to measure the knowledge base/competence pertinent to a particular clinical role, NPD specialists analyze performance expectations for the role, consult with subject matter experts (Toth, 2011), and construct tests of sufficient length to help assure accurate measurement. These processes help NPD specialists to represent practice accurately, increasing the validity of the measurement.

The world of measurement uses the term validity to describe the degree to which a measurement actually measures the intended characteristic (Bannigan & Watson, 2009). The credentialing world considers the related concepts of integrity, authenticity, and fidelity to explore how well certification, recertification, and other credentialing processes assure competence and continuing competence.

To measure competence accurately, multiple-choice items must avoid threats to validity. This article exposes some common flaws in multiple-choice items that interfere with accurate measurement and suggests remedies in the form of six tips to improve test items (see Figure 1).

Tip #1: Create a Practice Context

Place the question in the practice context to set the stage for nursing action. A solid practice context supports validity, but it is important to limit the length of the contextual story.

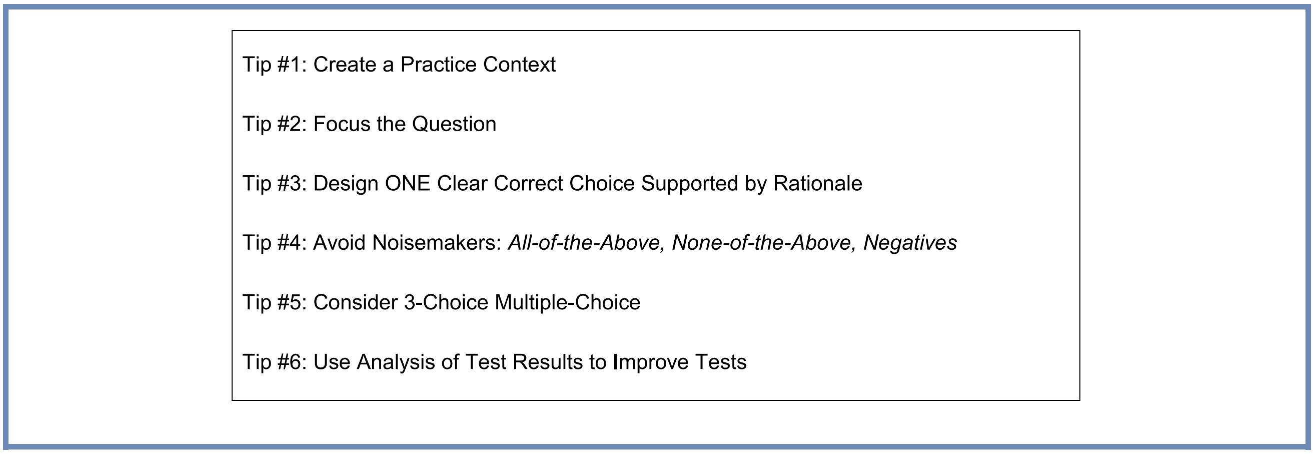
Karen Siroky, MSN, RN-BC, is Senior Clinical Director, AMN Healthcare, San Diego, California.

Bette Case Di Leonardi, PhD, RN-BC, is Independent Consultant in Education and Competency Management, Chicago, Illinois.

The authors have disclosed that they have no significant relationship with, or financial interest in, any commercial companies pertaining to this article.

ADDRESS FOR CORRESPONDENCE: Bette Case Di Leonardi, PhD, RN-BC, 56 West Schiller Street, Chicago, IL (e-mail: casedileonardi@gmail.com).

DOI: 10.1097/NND.0000000000000123



Tip #1: Create a Practice Context

Tip #2: Focus the Question

Tip #3: Design ONE Clear Correct Choice Supported by Rationale

Tip #4: Avoid Noisemakers: *All-of-the-Above, None-of-the-Above, Negatives*

Tip #5: Consider 3-Choice Multiple-Choice

Tip #6: Use Analysis of Test Results to Improve Tests

FIGURE 1 Six tips to improve test items.

Too much story adds reading load. Reading load refers to text in excess of what is needed to clearly express and measure as intended. Too little fails to set the stage for nursing action. At times, it may be acceptable to include a small amount of irrelevant information if the item is intended to measure the ability to sort out the significant findings. The test taker must use the information in the stem to answer the item. It is inappropriate to develop a situation involving a patient who has diabetes and then ask the question, "What is the normal range for fasting blood glucose?" A better use of the situation is to ask the test taker to analyze information presented about signs and symptoms and determine a course of action (see example in Figure 2).

When framing situations for test items, think through "What does the nurse do?" in the situation. Whenever possible, begin each option with a verb that states what the nurse does. The competent nurse does more than recognize an abnormal lab value. In fact, most clinical settings include reference ranges with lab reports. And so, the nurse must focus on patterns in findings, recognize why certain values are important, and decide what to do about abnormal findings. For example, instead of asking a question about the usual platelet count for a leukemia patient, ask why this value is important and what the nurse does about it. The correct answer is not a lab value but might rather be *increased risk for bleeding, implement nursing orders/standards to prevent injury*.

The patient teaching context might also provide a practice application of a fact or principle, for example, explaining how a pacemaker works or the purpose of oxygen therapy for a patient who has had a myocardial infarction (see example in Figure 3). However, putting facts and principles into lay language for a patient teaching test item may create options of unwieldy length. To create more succinct options, phrase the stem "You will explain in terms understandable to him that:"

Item writers often find it very easy to write items that test facts and principles. But to make a valid connection between knowledge/competence and practice/competency, the item writer must answer the question, "How does the nurse use this fact or principle in making a judgment?" The answer will suggest an item written at a higher cognitive level. Raise the bar by asking the test taker to exercise judgment, not simply recall a fact. An item that asks the test taker to interpret information provided and choose what action to take will usually be a higher cognitive level item, unless the test taker knows the correct answer because it is a familiar protocol and not a matter of professional judgment.

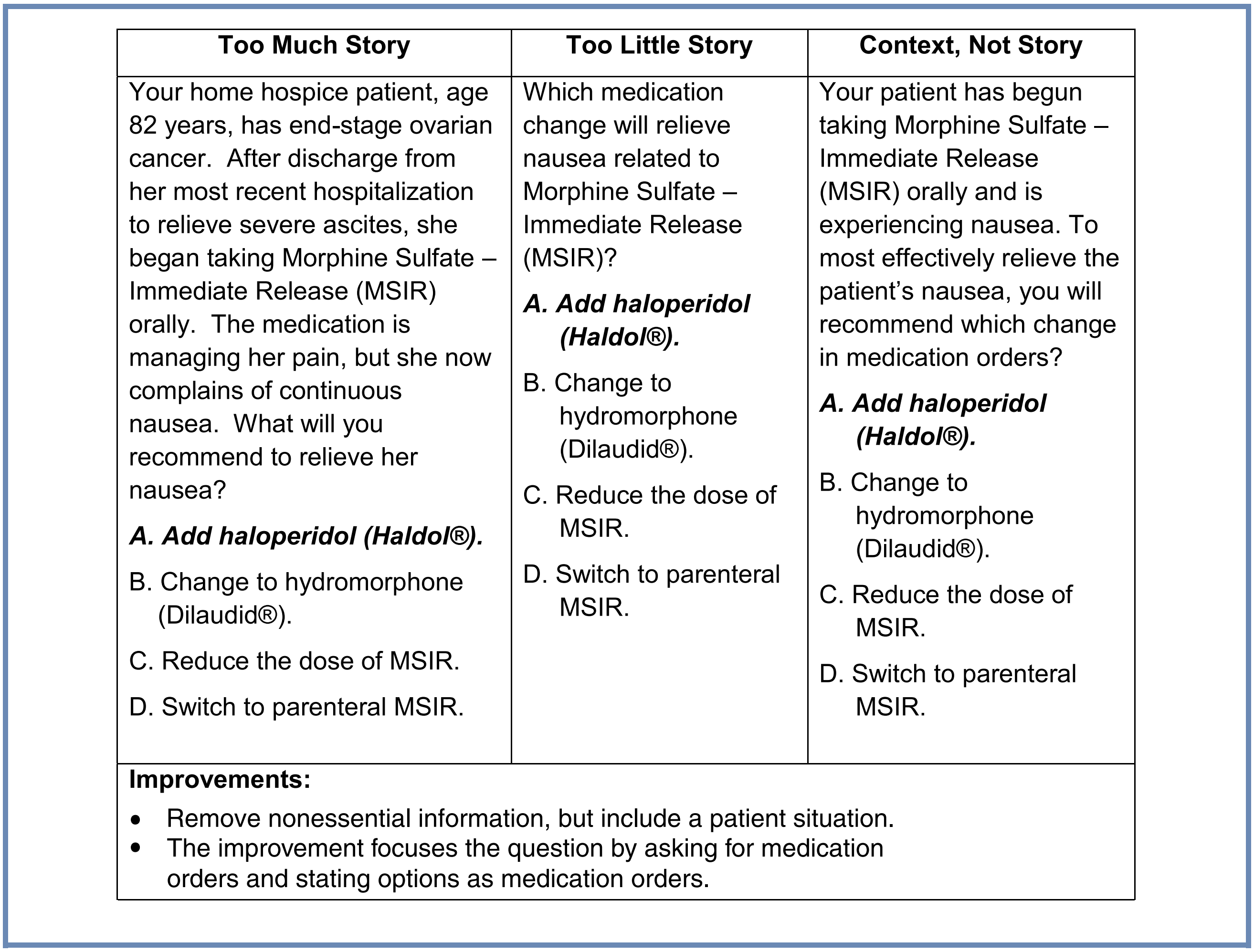
Remember to keep the practice context in focus by using common clinical mistakes and misunderstandings as distractors (incorrect options). Common mistakes make plausible distractors and may help to prevent mistakes when a test taker receives feedback on test performance. Avoid humorous or nonsensical distractors. Humor and nonsense may insult and distract the serious test taker. Meaningless distractors waste an opportunity to measure because the test taker will easily rule them out.

Tip #2: Focus the Question

A well-written stem poses a question or makes an incomplete statement. A well-written stem, and not the options, contains the central idea (Haladyna et al., 2002). Too much verbiage interferes with validity, because it detracts from the central point and creates reading load. The first words of the stem set the context, such as the patient, the situation, and the test taker's role. The last words tell the test taker what to look for in the options. For example, conclude a calculation item with "You will administer how many milliliters?" followed by options, each of which is a number of milliliters (see example in Figure 4).

Tip #3: Design One Clear Correct Choice Supported by Rationale

Sometimes test takers can successfully defend an answer other than the intended correct answer. Prevent this situation by locating current evidence-based rationale to support



Too Much Story	Too Little Story	Context, Not Story
Your home hospice patient, age 82 years, has end-stage ovarian cancer. After discharge from her most recent hospitalization to relieve severe ascites, she began taking Morphine Sulfate – Immediate Release (MSIR) orally. The medication is managing her pain, but she now complains of continuous nausea. What will you recommend to relieve her nausea? ***A. Add haloperidol (Haldol®).*** B. Change to hydromorphone (Dilaudid®). C. Reduce the dose of MSIR. D. Switch to parenteral MSIR.	Which medication change will relieve nausea related to Morphine Sulfate – Immediate Release (MSIR)? ***A. Add haloperidol (Haldol®).*** B. Change to hydromorphone (Dilaudid®). C. Reduce the dose of MSIR. D. Switch to parenteral MSIR.	Your patient has begun taking Morphine Sulfate – Immediate Release (MSIR) orally and is experiencing nausea. To most effectively relieve the patient's nausea, you will recommend which change in medication orders? ***A. Add haloperidol (Haldol®).*** B. Change to hydromorphone (Dilaudid®). C. Reduce the dose of MSIR. D. Switch to parenteral MSIR.
Improvements: • Remove nonessential information, but include a patient situation. • The improvement focuses the question by asking for medication orders and stating options as medication orders.		

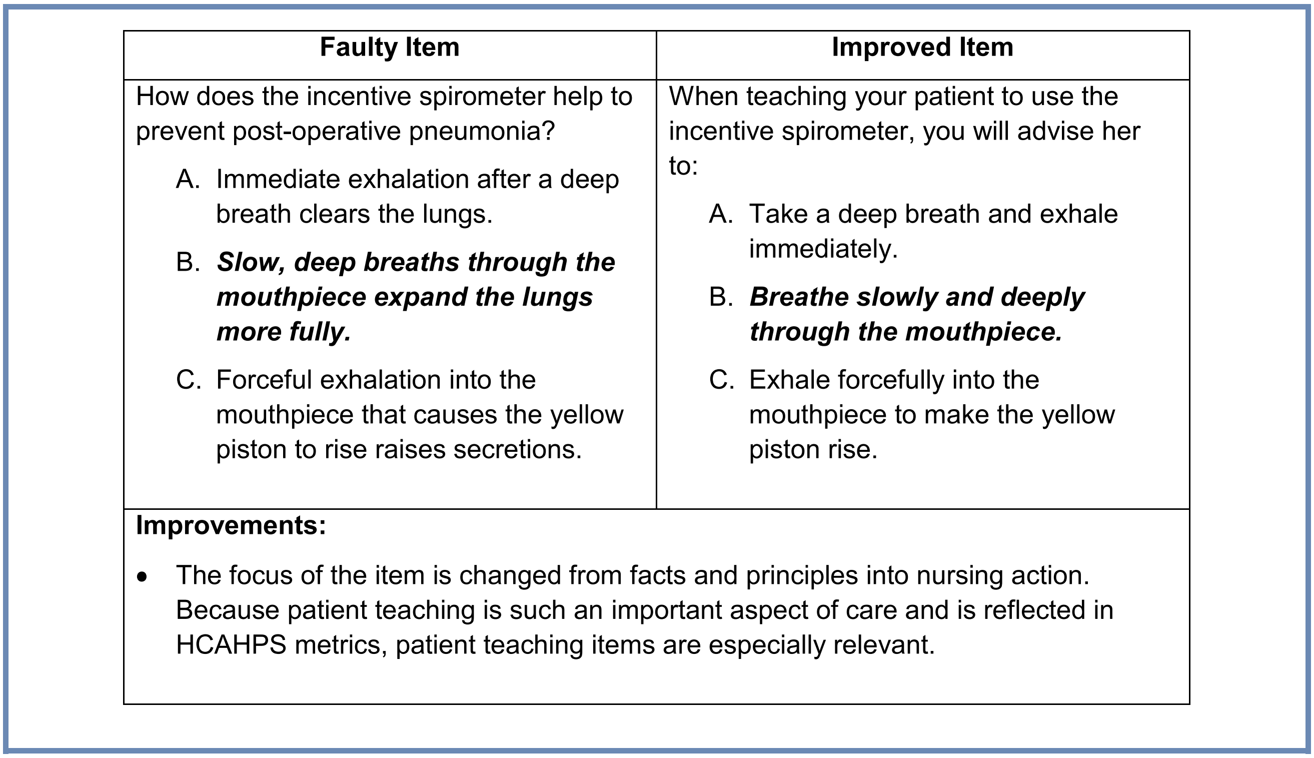
FIGURE 2 Tip #1: Create a practice context, not a story.

the correct option and the incorrectness of distractors. Doing so may lead to refining the options. The rationale and citation serve as learning resources for test takers.

Tip #4: Avoid Noisemakers: *All-of-the-Above, None-of-the-Above, Negatives*

All-of-the-above and none-of-the-above do not fit grammatically as the answer to a question or as a phrase to complete an incomplete sentence. If the item asks what action the nurse will take, all-of-the-above is not an answer to that question. In addition, when the test taker knows that more than one of four options are correct, he knows that all-of-the-above is the only possibility. Conversely, if he knows that one of the options is incorrect, he will rule out all-of-the above. In either case, the use of all-of-the-above as an option has made one of the incorrect options useless as a distractor for technical reasons that have nothing to do with measuring knowledge. Test takers may gravitate to the all-of-the-above option when they do not know the answer, figuring that it is a good guess. This is especially likely with test takers who have had plenty of previous experience with all-of-the-above as the correct answer.

As an alternative, create succinct, two- or three-part options in each distractor. If more parts are essential, place the one or two that everyone knows in the stem. For example, an item might test the knowledge of morphine, oxygen, nitroglycerin, and aspirin (MONA) as interventions to treat myocardial infarction. To decrease reading load for the test taker and eliminate for the item writer the challenge of coming up with three incorrect four-part options, the item writer might place one or more parts of MONA in the stem. For example, an item might read "For the patient who is experiencing a myocardial infarction (MI), immediate interventions include aspirin and:" Because the use of aspirin to treat MI is widely publicized, most test takers probably know that aspirin is correct and would choose only an option that contained aspirin. As another example, actions that apply in most situations such as "follow policy and procedure" or "document your observations" might be included in the stem. Offering such obvious correct answers does not help the test writer sort



Faulty Item	Improved Item
How does the incentive spirometer help to prevent post-operative pneumonia? A. Immediate exhalation after a deep breath clears the lungs. B. ***Slow, deep breaths through the mouthpiece expand the lungs more fully.*** C. Forceful exhalation into the mouthpiece that causes the yellow piston to rise raises secretions.	When teaching your patient to use the incentive spirometer, you will advise her to: A. Take a deep breath and exhale immediately. B. ***Breathe slowly and deeply through the mouthpiece.*** C. Exhale forcefully into the mouthpiece to make the yellow piston rise.
Improvements: • The focus of the item is changed from facts and principles into nursing action. Because patient teaching is such an important aspect of care and is reflected in HCAHPS metrics, patient teaching items are especially relevant.	

FIGURE 3 Tip #1: Create a practice context, raise the bar.

those who know the material from those who do not. The test development term for this sorting is discrimination.

Negatives of all kinds (such as none-of-the-above, double negatives in the item, all except, not) create noise. Test takers, especially if anxious or hurried, often misread negatives as positives and so answer incorrectly. Generally, it is more important to focus the test taker's attention on the correct action, rather than the incorrect action (see example in Figure 5). In addition, a negative requires a more complex thought process, which introduces noise and distracts from measuring what the item was intended to measure. Some recommend use of none-of-the-above as an option in calculation questions, in which the test taker must first perform the calculation before searching the options for the correct answer (McDonald, 2013).

Tip #5: Consider Three-Choice Multiple-Choice

Nursing examinations such as the licensing examinations and specialty certification examinations consist largely of four-choice multiple-choice items. NCLEX-RN® includes some alternative item types. Some certification examinations have introduced other formats. Academic programs also use alternatives to multiple-choice test items. However, the four-choice multiple-choice item predominates.

The literature (Edwards, Arthur, & Bruce, 2012; Rodriguez, 2005; Tarrant & Ware, 2012) suggests that the use of three-choice rather than four-choice multiple-choice items detracts little from validity and reliability and has decided advantages. It eliminates the difficulty of creating a fourth plausible option, greatly increasing efficiency of test development. Often item analysis reveals that, in a four-choice item, few or no test takers select one particular option. With less reading load per item, test takers can respond to three-choice items more quickly. Therefore, the test can present more items in the same time period. A longer test, one with more items, offers the advantage of increasing validity and reliability.

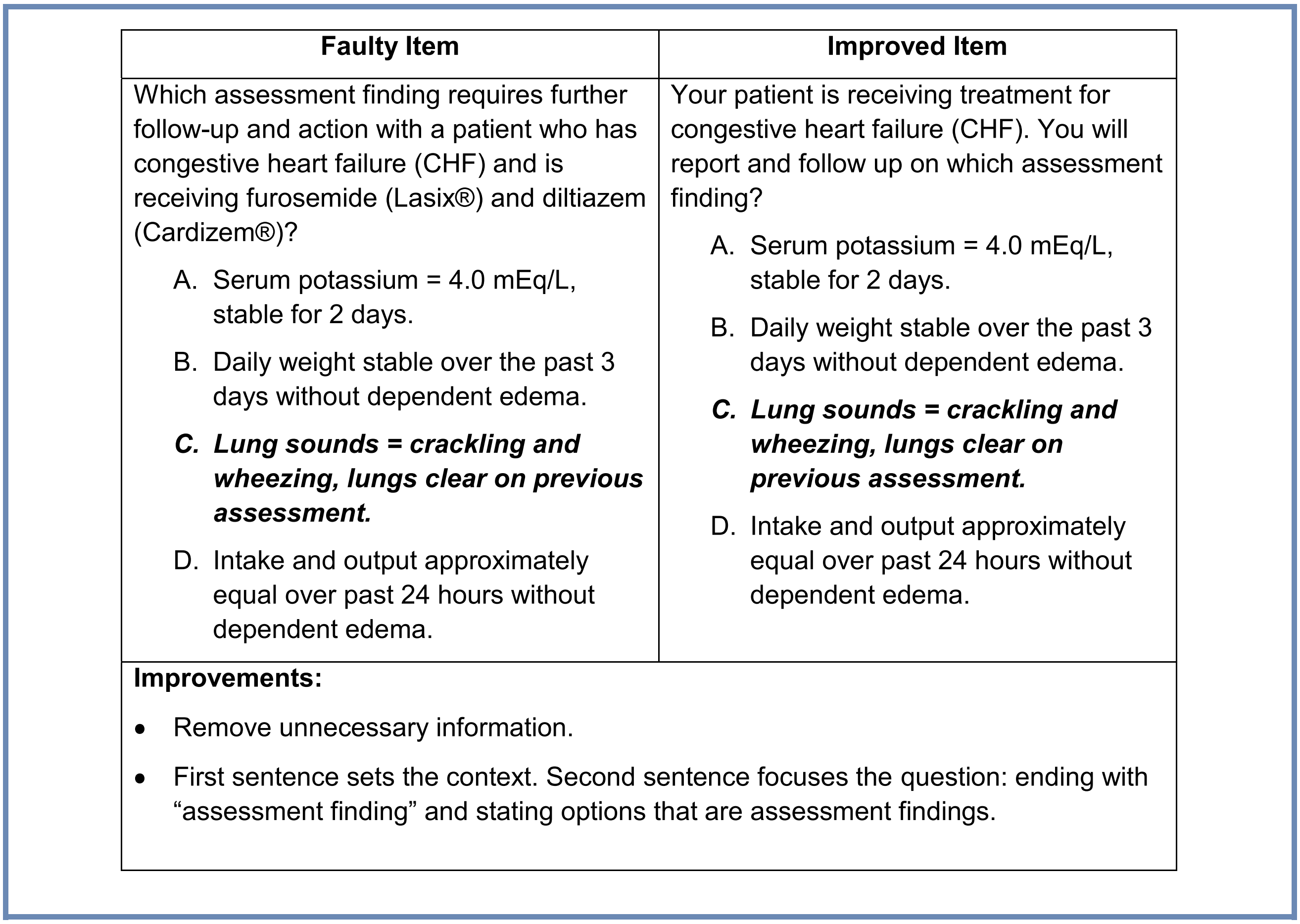
Tip #6: Use Analysis of Test Results to Improve Tests

Test items need regular updating to stay aligned with current evidence-based practice. In addition, technical improvements guided by analysis of test results strengthen validity. Four aspects of analysis of test results are especially useful:

- pass rate,
- difficulty,
- discrimination, and
- distractor analysis.

Although NPD specialists might welcome strict rules about using analysis of test results, they cannot escape the need to apply professional judgment in using the analysis. The analyzed data provide a source of knowledge, not a strict rule. Effective use of analysis of results requires the wisdom of professional judgment. Analysis of results tells test developers what to investigate, not what to do.

The discussion presented here is simplified to provide insight into the use of results. Most knowledge/competence



Faulty Item	Improved Item
Which assessment finding requires further follow-up and action with a patient who has congestive heart failure (CHF) and is receiving furosemide (Lasix®) and diltiazem (Cardizem®)? A. Serum potassium = 4.0 mEq/L, stable for 2 days. B. Daily weight stable over the past 3 days without dependent edema. ***C. Lung sounds = crackling and wheezing, lungs clear on previous assessment.*** D. Intake and output approximately equal over past 24 hours without dependent edema.	Your patient is receiving treatment for congestive heart failure (CHF). You will report and follow up on which assessment finding? A. Serum potassium = 4.0 mEq/L, stable for 2 days. B. Daily weight stable over the past 3 days without dependent edema. ***C. Lung sounds = crackling and wheezing, lungs clear on previous assessment.*** D. Intake and output approximately equal over past 24 hours without dependent edema.
Improvements: • Remove unnecessary information. • First sentence sets the context. Second sentence focuses the question: ending with “assessment finding” and stating options that are assessment findings.	

FIGURE 4 Tip #2: Focus the question.

assessment tests and continuing education course posttests that NPD specialists create are mastery tests. In mastery testing, the expectation is that most test takers will pass the test by obtaining a predetermined minimum passing score. A frequency distribution of scores is heavily skewed toward higher scores. Some of the published guidelines for interpretation of analysis of results apply to test results that conform to a bell curve rather than results with many high scores and few scores below the passing standard. Mastery testing usually yields results of a high pass rate and many items answered correctly by most test takers. This situation influences the statistics used in analysis of results. For complete discussion and more precise information about computation, see McDonald (2013) and Oermann and Gaberson (2013).

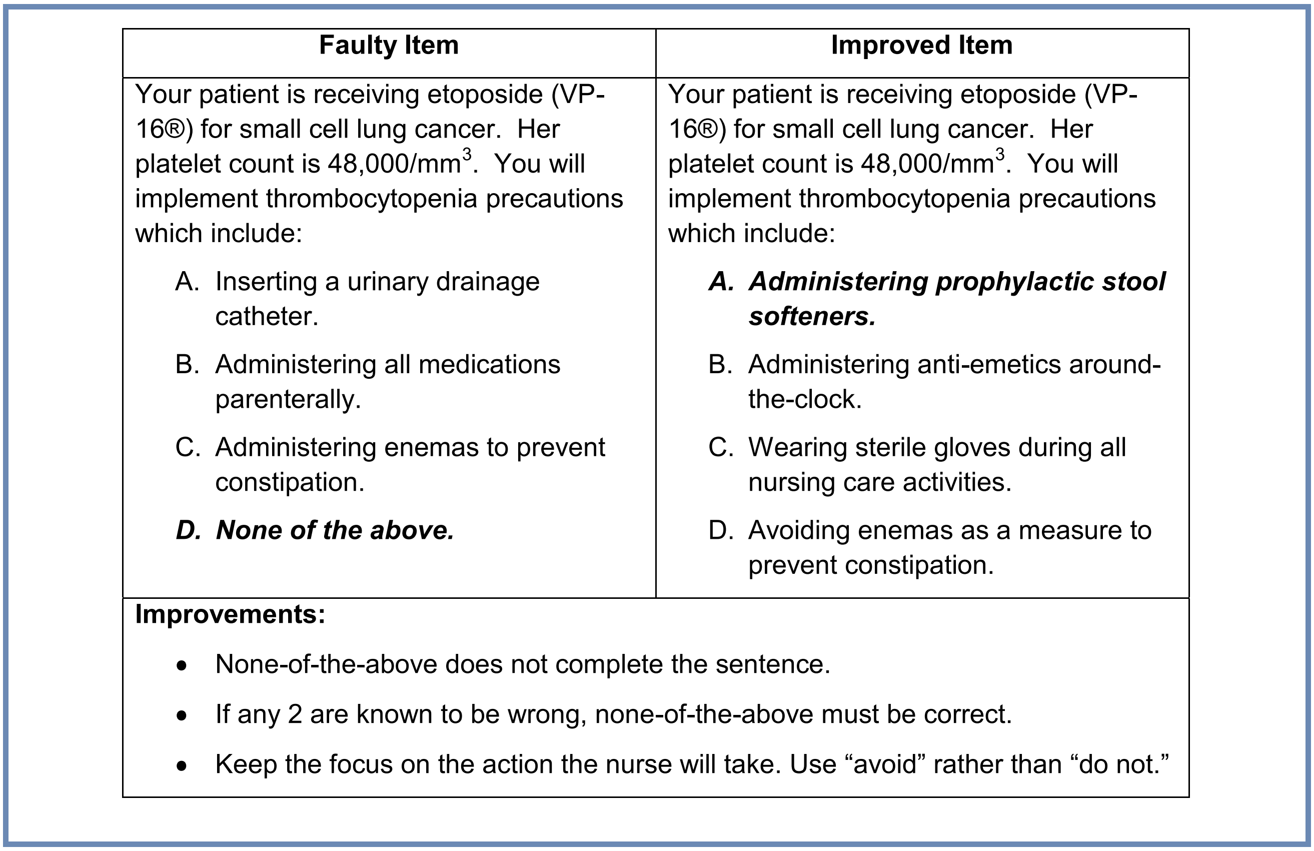
Pass rate equals the percentage of test takers who passed the test. A discussion of methods for setting a passing or cutoff score is beyond the scope of this article. Because NPD testing is competence and safety related, tests used in NPD often require a percentage correct of at least 80% and occasionally 100%. In NPD, test takers who do not pass may receive remediation to assure that they know the correct answer. The remediation process may reveal faults in a particular test item, such as ambiguity or perhaps two correct answers to an item intended to have only one correct answer.

In continuing education posttesting, educators may tolerate a lower pass rate. Sometimes participants take the posttest more than once to obtain passing scores, but their initial scores may be included in the pass rate calculation.

If the pass rate is 100%, one may question whether the test may be too easy. Perhaps some items need to present greater challenge. Perhaps the topics are too basic. Perhaps 100% is essential because of assure safe practice. Similarly, a low pass rate requires investigation.

Difficulty equals the percentage of test takers who answered an item correctly. Difficulty may be calculated for each item and for the test overall as an average of all the individual item difficulties. Paradoxically, 100% or 1.0 difficulty means that all test takers answered correctly: the higher the difficulty value, the easier the item for this group of test takers. For example, if difficulty = 0.75, 75% of test takers answered correctly.

As a useful rule of thumb, investigate any item answered correctly by fewer than 75% of test takers. Investigate does not dictate whether to revise, eliminate, or retain. It simply



Faulty Item	Improved Item
Your patient is receiving etoposide (VP-16®) for small cell lung cancer. Her platelet count is 48,000/mm^3. You will implement thrombocytopenia precautions which include:	Your patient is receiving etoposide (VP-16®) for small cell lung cancer. Her platelet count is 48,000/mm^3. You will implement thrombocytopenia precautions which include:
A. Inserting a urinary drainage catheter. B. Administering all medications parenterally. C. Administering enemas to prevent constipation. ***D. None of the above.***	***A. Administering prophylactic stool softeners.*** B. Administering anti-emetics around-the-clock. C. Wearing sterile gloves during all nursing care activities. D. Avoiding enemas as a measure to prevent constipation.

Improvements:

- None-of-the-above does not complete the sentence.
- If any 2 are known to be wrong, none-of-the-above must be correct.
- Keep the focus on the action the nurse will take. Use “avoid” rather than “do not.”

FIGURE 5 Tip #4: Avoid noisemakers: *all-of-the-above, none-of-the-above, negatives.*

means to use professional judgment in exploring the poor performance and adjust the item, the learning experience, or simply enforce the expectation.

Discrimination equals the difference between the number of high scorers/passers who answered an item correctly and the number of low scorers/nonpassers who answered an item correctly. The desired result is that as many or more high scorers/passers answered correctly than did low scorers/non-passers. When more low scorers/non-passers than high scorers/passers answer correctly, it suggests that something is amiss with the item. This situation is called negative discrimination, because the number of high scorers/passers minus the number of low scorers/non-passers yields a negative result. Perhaps the item is ambiguous, or perhaps advanced knowledge leads a test taker away from the intended correct answer.

Distractor analysis equals the number and status (high or low scorers) of test takers who choose each incorrect option. As noted previously, distractors must be plausible and should attract test takers who do not know the correct answer. When few or no test takers choose a particular distractor, it suggests a need to make the distractor more challenging. Related to discrimination, it signals a problem with a distractor if low scorers answer correctly but more high scorers choose a particular distractor. For NPD purposes, computation of distractor analysis is rarely indicated. However, it is useful to see whether test takers are choosing distractors and if distractors might be improved.

When most test takers pass, there will be many distractors that are chosen by few or no test takers. Nevertheless, the test developer needs to remain alert for opportunities to improve distractors.

A number of commercial software and Web-based programs are available to assist with analysis of results. In addition, Internet sources explain how to create spreadsheet formulae to analyze test results. With an understanding of the meaning and significance of these four aspects of analysis of results, the NPD professional can perform a simple analysis of at least selected items, even without a sophisticated program.

CONCLUSION

Careful planning and attention to the tips this article presents contribute to item validity. But carelessness can still sabotage accurate measurement. Take the final important steps and carefully proofread, edit, format, and correct errors in grammar, punctuation, capitalization, and spelling (Haladyna et al., 2002).

Valid, effective test items are essential tools in the NPD specialist’s competence assessment/validation tool kit. Valid measurement builds credibility and confidence in the NPD specialist’s expertise in documenting competence, both for stakeholders in the organization and for the clinicians who take these tests. Strengthening test development skills aids the NPD specialist in demonstrating the value of NPD in the organization's competency management model.

References

Bannigan, K., & Watson, R. (2009). Reliability and validity in a nutshell. *Journal of Clinical Nursing*, *18*, 3237–3243.

Edwards, B. D., Arthur, W., & Bruce, L. (2012). The three-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*(1), 65–81.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*, 309–334.

McDonald, M. E. (2013). *The nurse educator's guide to assessing learning outcomes* (3rd ed.). Burlington, MA: Jones & Bartlett Learning.

Oermann, M., & Gaberson, K. (2013). *Evaluation and testing in nursing education* (4th ed.). New York, NY: Springer Publishing Company.

Rodriguez, M. (2005). Three options are optimal for multiple-choice items: A meta- analysis of 80 years of research. In *Educational Measurement: Issues and Practice*, *24*(2),3–13.

Sutherland, K., Schwartz, J., & Dickison, P. (2012). Best practices for writing test items. *Journal of Nursing Regulation*, *3*(2), 35–39.

Tarrant, M., & Ware, J. (2012). A comparison of the psychometric properties of three- and four-choice multiple-choice questions in nursing assessments. *Nurse Education Today*, *30*, 539–543.

Toth, J. (2011). Assessment tool for medical-surgical nursing (MED-SURG BKAT)© and implications for in-service educators and managers. *Nursing Forum*, *46*(2), 110–116.

Notice: Online CE Testing Only Coming in 2015!

Starting with the first issue of 2015, the tests for CE articles will appear only in the online version of the issue, and all tests must be completed online at www.nursingcenter.com/ce/JNSD. Simply select the CE article you are interested in. Both the article and the test are available there. You will no longer have the option to mail or fax in the test.

If you haven't done so already, you will want to create a user account for yourself in Nursing Center's CEConnection - it's free to do so! Look for the Login link in the upper right hand corner of the screen.

For more than 22 additional continuing education articles related to professional development, go to NursingCenter.com\CE.