Tests, Quizzes, and Self-Assessments: How to Construct a High-Quality Examination

Anouk van der Gijp[1]
Cécile J. Ravesloot[2]
Olle Th. J. ten Cate[3]
Jan P. J. van Schaik[2]
Emily M. Webb[1]
David M. Naeger[1]

OBJECTIVE. The purposes of this article are to highlight aspects of tests that increase or decrease their effectiveness and to provide guidelines for constructing high-quality tests in radiology.

CONCLUSION. Many radiologists help construct tests for a variety of purposes. Only well-constructed tests can provide reliable and valuable information about the test taker.

Keywords: continuing medical education, medical education, radiology assessment, radiology education, test quality

DOI:10.2214/AJR.15.15944

Received November 25, 2015; accepted after revision February 5, 2016.

[1]Department of Radiology and Biomedical Imaging, University of California, San Francisco, 505 Parnassus Ave, M-391, San Francisco, CA 94143-0628. Address correspondence to D. M. Naeger (david.naeger@ucsf.edu).

[2]Department of Radiology, University Medical Center Utrecht, Utrecht, The Netherlands.

[3]Center for Research and Development of Education, University Medical Center Utrecht, Utrecht, The Netherlands.

This article is available for credit.

AJR 2016; 207:339–343

0361–803X/16/2072–339

Tests, Tests, and More Tests

By the time radiologists begin to practice, they have taken hundreds of tests. There are high-stakes tests, such as the American Board of Radiology examinations and the United States Medical Licensing Examination. And there are lower-stakes tests, such as the short quizzes following training or compliance modules, which have become commonplace in many hospitals. In addition to taking tests, many radiologists create tests. Some serve on national, local, or institutional committees responsible for learner assessment. Even those who are not primarily engaged in radiology education may be confronted by test construction tasks in other settings, as in committees reviewing compliance with hospital initiatives (e.g., testing knowledge about patient safety initiatives, such as IV contrast administration policies and policies regarding MRI zones) or related to their research activities (e.g., testing of study subjects). Though testing is commonplace and important, it is worth pointing out that tests should generally be one part of the overall assessment of learners.

Why Tests Must Be of High Quality

Understanding what makes an effective test is important. Test takers want tests that assess essential knowledge or skills rather than irrelevant knowledge. Examinees want questions that match their expectations and are perceived as fair. A well-constructed examination gives learners confidence in the value and meaning of their test results and emphasizes the importance of the underlying subject.

For test makers, multiple steps can be followed to ensure that a learner's knowledge or skills are accurately evaluated. Simply asking a question does not ensure appropriate testing. A poorly constructed test can even interfere with the effective demonstration of knowledge or skills. Alternatively, learners may score higher or lower than is warranted if the examination questions do not adequately discriminate between learners who have and have not learned the important material. Especially in the field of medical education, low-quality tests can actually do harm, because patient safety may be at risk when learners pass examinations that they rightfully should have failed. As radiologists, we have a responsibility to train and deliver competent physicians who meet level-appropriate standards. This requires high-quality tests all along the training pathway. In this article, we describe what constitutes a high-quality test and outline seven steps for designing a high-quality test.

What Constitutes High Quality?

There are three fundamental components of a high quality test: validity, reliability, and fairness [1] (Table 1).

Validity

Validity is the property by which a test actually measures what it is intended to measure. Those responsible for test construction must ensure that the assessment will result in an accurate judgment about the test taker's knowledge or skills and that decisions based on test scores are appropriate for the purpose of the test [2].

TABLE 1: Fundamental Aspects of Test Quality

Quality Aspect	What Does It Mean?	What Is the Goal?
Validity	Whether a test measures what it is intended to measure	A test should measure what is intended, tailored to its specific purpose.
Reliability	Precision and reproducibility of test scores	Test scores should be as consistent as reasonably achievable, especially for high-stakes tests.
Fairness	Accessibility of testing	A test should not advantage or disadvantage anyone on the basis of characteristics that are irrelevant to the tested capabilities.

The first component of validity is the subject matter of the test. Specifically, the content of the test must match the purpose of the test [1, 3]. For example, a test on chest radiograph interpretation comprising only lung nodule cases would not assess a reader's interpretation of acute lung disease. Such a test would therefore have limited predictive value for image interpretation performance in the emergency department.

The level of testing can also fail to match the goal of the test. Professionals perform work at various levels. In radiology, for example, detecting lesions on images and generating differential diagnoses require higher levels of performance than does remembering facts. A test requiring factual radiologic knowledge recall alone would not assess the skills required to detect imaging abnormalities and generate a differential diagnosis. Various frameworks, such as the Bloom taxonomy [4] and the Miller pyramid [5], can be used to determine the appropriate level of assessment. In this review, we use Miller's framework because it specifically describes different levels for assessing clinical competence. The details of this framework are discussed later (see Constructing a High-Quality Test: Seven Steps to Success).

Reliability

Reliability refers to the consistency of scores across repeat administrations of a test [1]. A reliable test should yield reproducible scores, or at least a similar ranking of participants. To determine reliability, one often does not have the opportunity to administer the same test repeatedly. However, breaking up a larger test into two or more parts with comparable questions and correlating the scores can provide information about the overall reliability; parts should perform similarly if they test the same content.

Reliability generally improves when a test contains more questions, analogous to including more patients (i.e., a larger *n*) in research studies. Examiners also should realize that anything external to the test content that interferes with a test taker's thinking process may decrease the reliability of the answers, such as conditions in the examination room and personal factors like having a headache.

When a test taker scores near the pass-fail border, the importance of reliability increases [1]. If a test is unreliable, chance alone can alter whether a learner passes or fails. Some tests, such as oral tests, are notoriously unreliable: graders tend to vary between administrations, and their assessments can vary among one another. Improving reliability was one of the goals of the new American Board of Radiology qualifying examination structure, which replaces oral examinations with computer-based tests [6, 7].

Fairness

A fair test treats all subjects equally and does not advantage or disadvantage anyone on the basis of characteristics that are irrelevant for the capabilities being tested [1]. For example, if language ability is not being tested, the test should be in a language that has been mastered by the participant. For tests of medical professionals, however, mastery of the language predominantly used by the medical community may be part of the test. Tests involving medical terminology require extra attention, because the use of terminology tends to vary among physicians and institutions and may change over time. For example, the term "Wegener granulomatosis" has recently been replaced by "granulomatosis with polyangiitis." A fair test should not penalize the test taker for not knowing the new name, unless knowledge of the transition is part of the test goal.

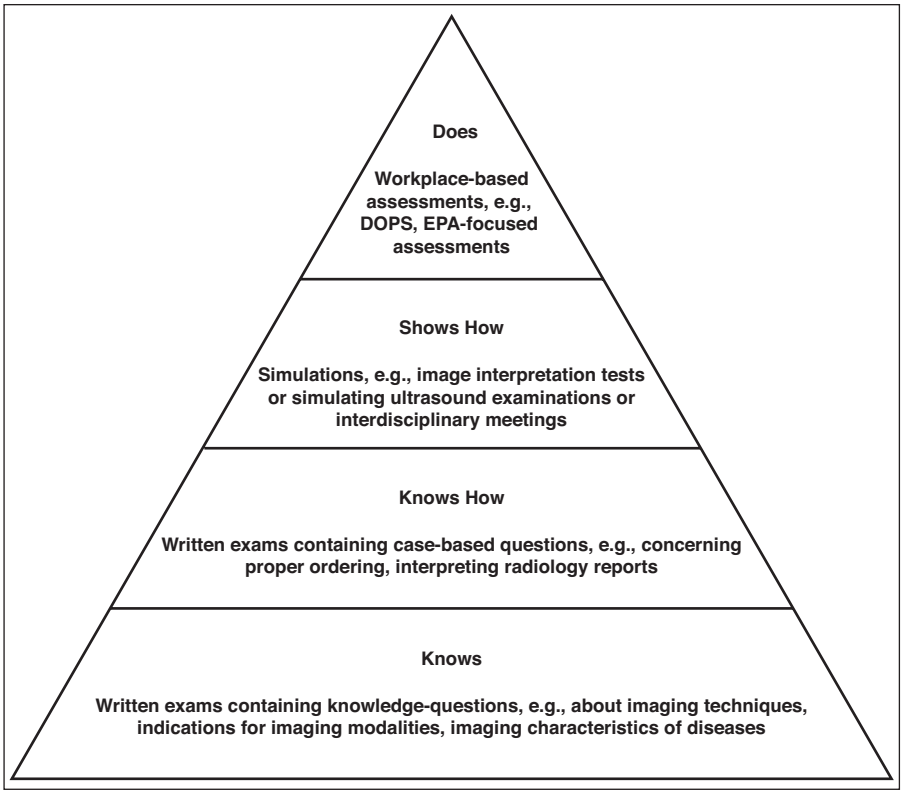

Fig. 1—Chart shows testing methods for different levels of radiology tests. DOPS = direct observation of procedural skills, EPA = entrustable professional activities.

Constructing a High-Quality Test: Seven Steps to Success

Creating a valid, reliable, and fair examination requires effort. The following seven discrete steps can help direct the process. Each step can be increasingly applied as test creators become more experienced.

1. Identify the Purpose and Content of the Test

Before deciding what to test, educators must decide what must be taught [8, 9]. This seems obvious but is often overlooked in education. In medical school, for example, one may decide that noninterpretative skills, such as proper ordering, are more important than interpretative skills [10]. For radiology residents, one may want to teach how to detect abnormalities (teaching how to detect is often underemphasized) and which diseases each finding can represent. Regardless of the field, reflecting on what must be learned is essential. What to teach can be informally determined by the educator or can be formally stated in learning objectives [11]. The content of the test should reflect the learning objectives of the educational activities, or part of them, when the test is part of an assessment program with multiple tests.

After what should be taught is decided, the test must be matched to the learning objectives. The Miller pyramid is a useful framework for deciding how to assess clinical skills [5]. This framework delineates four levels of assessment: knows, knows how, shows how, and does (Fig. 1). For the lowest level, knows, test questions can be constructed without a specific clinical context. For example, a straightforward multiple-choice question can ask for the signal characteristics of blood on MR images. A context-rich format (e.g., clinical vignette) is needed for the knows how level, which measures the ability to apply knowledge, such as decision making in ordering studies based on certain symptoms. Assessment on the shows how level is best conducted with simulation scenarios that approach reality (e.g., interpreting and making reports on selected images at a PACS workstation or performing ultrasound on a phantom). If a PACS workstation is unavailable, computerized testing with stack images can improve testing of image interpretation skills [12, 13]. The highest level of assessment, does, requires observation of the learner in clinical practice (e.g., evaluating preliminary image readings by residents on call), in which cases are not preselected and interpretation takes place in a busy environment. Many assessment instruments are developed for this purpose, including direct observation of procedural skills [14] and case-based discussions related to entrustable professional activities [15].

The ultimate goal of the test requires substantial consideration. A test can have many different goals [3], such as to graduate students, to measure their ability to take care of patients, or to provide learners with feedback about their progress. The purpose of the test should guide both test construction and pass-fail decisions.

2. Create a Test Blueprint

A test blueprint is a template that details the specific content of a test [16]. It is typically designed as a matrix listing the number of questions per learning objective. Questions can also be categorized by content (e.g., radiology modalities) or level of assessment (e.g., Miller levels). There are many variations of test matrices, which can also include question characteristics, such as question types, symptoms, and image modalities. The purpose of the test blueprint is to ensure that the test will cover all topics, question types, and modalities intended, preventing overrepresentation and underrepresentation. In the construction of multiple test versions, a test blueprint helps ensure the versions are roughly equivalent. The test blueprint should represent generally important (e.g., life-threatening conditions) and commonly encountered topics (e.g., common diseases) more than rare or less important conditions [8]. Figure 2 shows a simple example blueprint.

3. Assemble a Question Bank

Tests are composed of questions. Rather than starting by creating each entire test separately, it is often best to start writing questions and categorize them with labels based on the factors specified in the blueprint. Once the question bank is partially populated, specific questions can be created to fill gaps. No matter how well thought out the test purpose and test blueprint are, low-quality questions can ruin validity and reliability. The following steps are a guide to creating a good question.

Choose a response format that fits to the test goal—Closed-format (multiple choice) questions can compose reliable tests owing to the ability to have a large number of questions and objective scoring [17, 18]. Open-ended questions may be easier to construct, but the grading can be time-consuming with detailed answer keys required to ensure objectivity [3, 18]. To our knowledge, there is no evidence that open-ended questions test knowledge and skills on a higher level than closed-format questions [3, 19]; therefore, some authors [3, 18] recommend closed-ended questions. Closed-format question types include true-false questions, single-best-option multiple choice, and multiple-response questions. True-false questions mainly test recall of facts (knows level) [20]

Overarching Purpose

Test for competence in image utilization and basic interpretation skills of interns who will be managing emergency department patients

Learning Objectives

After the training, the intern is able to
Learning objective 1: list different sorts of radiologic study protocols and their indications
Learning objective 2: understand radiology reports in acute cases
Learning objective 3: detect abnormalities on radiographs in acute cases
Learning objective 4: diagnose prevalent or critical acute diseases on the basis of clinical history and imaging

Test Blueprint With Numbers of Items	Musculoskeletal Radiology	Chest Radiology	Abdominal Radiology	Neuroradiology	Question Type per Area		
					MC	CB	IB
Learning objective 1 (knows)	3	3	3	3	2	1	0
Learning objective 2 (knows how)	4	4	4	4	1	3	0
Learning objective 3 (shows how)	5	5	1	0	0	0	All
Learning objective 4 (does)	1	2	2	2	0	0	All

MC = multiple choice without clinical case, CB = case based, IB = image based.

Fig. 2—Example shows learning objectives and test blueprint.

TABLE 2: Frequently Used Parameters of Test Quality

Parameter	What Does It Tell?	How Does It Work?	Recommendations
Test level			
Cronbach α	Overall test reliability	Measures internal consistency of test	High level of consistency between questions is desirable, preferably > 0.8 (> 0.9 for high-stakes tests).
Question level			
p	Difficulty index	Shows proportion of students with correct answer	Moderately difficult questions are better than very difficult or very easy ones so they can contribute to discrimination.
Item-test correlation	Discrimination power index	Correlates question score variation with overall score variation	Highly discriminating questions are desirable, preferably > 0.30; negative values can be caused by flaws.
Answer option level			
Item response pattern	Answer distribution	Shows percentage of students who selected each answer option	Each answer choice should be selected occasionally to indicate it is a plausible distractor.

and are not easy to write without flaws. Single-best-option multiple choice and multiple-response questions are preferred for testing on higher levels [8]. Two advanced question types that can be particularly useful in image interpretation testing are hot-spot and long-menu questions. A hot-spot question involves marking a structure on an image, which can test perception of abnormalities and recognition of normal anatomy [12, 13]. A long-menu question is a variant of a multiple-choice question that includes long lists of answer options [21, 22]. If an examinee is required to generate a single diagnosis or a short differential diagnosis list, the sheer number of possibilities better simulates clinical practice and eliminates guessing.

Avoid question-writing flaws—Flaws can threaten reliability and validity. For example, a common question flaw is to use negatives such as "not" and "except" in the stem or in answer options, testing reverse thinking rather than knowledge. A flawed question may also give clues to the correct answer, as when the correct answer is much longer than other options or when incorrect answer options contain absolute terms such as "never" and "always." Teachers trained in item construction make fewer errors [23]. If formal training is not available, several informative publications on question flaws are available for review [8, 24–26]. Given the extensive resources published on question flaws, including in the radiology literature, we are not reviewing them in detail.

Select an adequate number of questions to reach a reasonable level of reliability—In general, approximately 100 well-written multiple choice items are needed to reach sufficient reliability of high-stakes tests [3]. Increasing the number of questions will improve reliability as long as the questions test the same knowledge and skills. Of course, there is a tradeoff with time constraints. Questions with images take more time to complete, especially when stacks of images are used [27].

4. Construct the Test

With a completed question bank and sufficient numbers of questions in each category specified in the blueprint, the test can be composed. The numbers of questions specified in the matrix of the blueprint define how many questions are to be pulled from each category. Multiple versions of a test can be created, if desired. Different tests can also be created from one question bank provided the questions meet the goals of each test constructed. For example, the Radiology Exam-Web question database can be used to create a core clerkship test or a test specific to an interventional radiology rotation [28].

5. Administer the Test

Testing conditions should be comfortable and reasonable. Noisy environments and uncomfortable temperatures should be avoided. Digital tests should be administered on fast computers with appropriate resolution displays [29]. In general, testing situations should not easily allow cheating. In addition to providing an appropriate environment, everyone taking the test should be provided as similar an environment as possible, for the sake of fairness. Image quality on each computer should be similar between test takers. Bringing one's own device is therefore problematic for fair image interpretation testing.

6. Score the Test and Interpret the Results

Important topics should contribute more substantially to final scores than less important topics. It is preferable to accomplish this by increasing the number of questions addressing important topics instead of weighting them more, because a larger sample increases score reliability [8]. A pass-fail interpretation of test results is often used to distinguish proficiency. A deliberate and accurate determination of passing score is essential for validity [1]. Passing scores can be based on test scores of the group (norm-referenced standard) or based on a predefined passing score (criterion-referenced standard). The method chosen depends on the purpose of the test. For tests of proficiency, passing scores relative to the group could be considered unfair because a proportion of the test takers will always fail regardless of absolute proficiency. For the purposes of admission into a school or career track, defining a proportion of the group to pass may be necessary owing to limited available spots. Different methods of establishing a passing score are explained in detail in the Association for Medical Education in Europe guide number 18 [30].

7. Perform Post Hoc Analysis

Post hoc analysis of the overall test quality and the value each question adds is important for quality assurance and test improvement. The most basic level of post hoc analysis is simply going over the test with the takers immediately after the examination. With an engaged group of test takers, a surprisingly large amount of information can be gleaned about which questions were too easy, too hard, confusing, or esoteric. For high-stakes examinations, a test and its questions should be evaluated in a more systematic and quantitative way. Specifically, the test as a whole can be evaluated, individual questions can be evaluated, and individual answer choices can be evaluated.

The reliability or internal consistency of the entire test can be assessed with the Cronbach alpha value [31]. This metric is used to evaluate the degree to which individual question scores correlate with each other. If a group of questions are testing the same body of knowledge and skills, test takers should generally score consistently on the questions in that group. Calculating Cronbach alpha generally requires a statistical software program.

Individual questions can be analyzed as to difficulty and discriminative power. Item difficulty is relatively intuitive: the more dif-

ficult a question, the fewer learners will respond correctly. However, a low number of correct responses does not necessarily mean that the content of the question is difficult. Questions may be difficult for a number of reasons, including poor or confusing wording, esoteric content, and a bad format. Discrimination is the degree to which learners with high overall test scores answer a particular question correctly and vice versa. Essentially, discriminating those who know the content from those who do not is a measure of the effectiveness of an item. Question discrimination and question difficulty are related because very difficult questions and very easy questions tend to be poor discriminators. If an item is so easy that almost every test taker answers it correctly or so difficult that almost all answer it incorrectly, then it is very difficult to discriminate who actually knows the content. Statistical measures of question difficulty and discrimination that can be calculated are listed in Table 2.

Response patterns for individual answer options can also be analyzed. For example, if an answer choice is not selected by any test taker, it is probably not a plausible distractor and should be replaced or omitted. Table 2 lists frequently used parameters for quantitative evaluation for each level. Statistical details are available in a comprehensive guide from the Association for Medical Education in Europe [32].

Conclusion

All radiologists take written or electronic tests, and many radiologists are involved in test construction. To yield reliable results, tests should be of high quality. This article summarizes key aspects of test quality and provides basic guidelines for constructing an efficient and high-quality test.

References

1. AERA, APA, NCME. *Foundations. standards for educational and psychological testing: Part 1.* Washington, DC: American Educational Research Association, 2014:11–72
2. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ* 2015; 49:560–575
3. Hawkins RE, Swanson DB. Using written examinations to assess medical knowledge and its application. In: Holmboe ES, Hawkins RE, eds. *Practical guide to the evaluation of clinical competence*. Philadelphia, PA: Mosby, Elsevier, 2008:42–59
4. Bloom BS. *Taxonomy of educational objectives: handbook I—cognitive domain*. New York, NY: David McKay, 1956
5. Miller GE. The assessment of clinical skills/competence/performance. *Acad Med* 1990; 65(suppl): S63–S67
6. Becker GJ, Bosma JL, Guiberteau MJ, Gerdeman AM, Frush DP, Borgstede JP. ABR examinations: the why, what, and how. *Radiology* 2013; 268:219–227
7. Hollingsworth CL, Wriston CC, Bisset GS, et al. American Board of Radiology certifying examination: oral versus computer-based format. *AJR* 2010; 195:820–824
8. Case SM, Swanson DB. *Constructing written test questions for the basic and clinical sciences*. Philadelphia, PA: National Board of Medical Examiners, 1998
9. Kondo KL, Swerdlow M. Medical student radiology curriculum: what skills do residency program directors believe are essential for medical students to attain? *Acad Radiol* 2013; 20:263–271
10. Naeger DM, Webb EM, Zimmerman L, Elicker BM. Strategies for incorporating radiology into early medical school curricula. *J Am Coll Radiol* 2014; 11:74–79
11. Webb EM, Naeger DM, Fulton TB, Straus CM. Learning objectives in radiology education: why you need them and how to write them. *Acad Radiol* 2013; 20:358–363
12. Ravesloot CJ, van der Gijp A, van der Schaaf MF, et al. Support for external validity of radiological anatomy tests using volumetric images. *Acad Radiol* 2015; 22:640–645
13. Ravesloot CJ, van der Schaaf MF, van Schaik JP, et al. Volumetric CT-images improve testing of radiological image interpretation skills. *Eur J Radiol* 2015; 84:856–861
14. Norcini J, Burch V. Workplace-based assessment as an educational tool: AMEE guide no. 31. *Med Teach* 2007; 29:855–871
15. Ten Cate O, Chen HC, Hoff RG, Peters H, Bok H, van der Schaaf M. Curriculum development for the workplace using entrustable professional activities (EPAs): AMEE guide no. 99. *Med Teach* 2015; 37:983–1002
16. Sales D, Sturrock A, Boursicot K, Dacre J. Blueprinting for clinical performance deficiencies: lessons and principles from the General Medical Council's fitness to practise procedures. *Med Teach* 2010; 32:e111–e114
17. Hift RJ. Should essays and other "open-ended"-type questions retain a place in written summative assessment in clinical medicine? *BMC Med Educ* 2014; 14:249
18. Schuwirth LW, van der Vleuten CP. Different written assessment methods: what can be said about their strengths and weaknesses? *Med Educ* 2004; 38:974–979
19. Palmer EJ, Devitt PG. Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? Research paper. *BMC Med Educ* 2007; 7:49
20. Schuwirth LW, Verheggen MM, van der Vleuten CP, Boshuizen HP, Dinant GJ. Do short cases elicit different thinking processes than factual knowledge questions do? *Med Educ* 2001; 35:348–356
21. Schuwirth LW, van der Vleuten CP, Stoffers HE, Peperkamp AG. Computerized long-menu questions as an alternative to open-ended questions in computerized assessment. *Med Educ* 1996; 30:50–55
22. Rotthoff T, Baehring T, Dicken HD, et al. Comparison between long-menu and open-ended questions in computerized medical assessments: a randomized controlled trial. *BMC Med Educ* 2006; 6:50
23. Webb EM, Phuong JS, Naeger DM. Does educator training or experience affect the quality of multiple-choice questions? *Acad Radiol* 2015; 22:1317–1322
24. DiSantis DJ, Ayoob AR, Williams LE. Prevalence of flawed multiple-choice questions in continuing medical education activities of major radiology journals. *AJR* 2015; 204:698–702
25. American Board of Radiology website. Item-writing guide. www.theabr.org. Accessed April 5, 2016
26. Haladyna TM, Downing SM, Rodriguez Adrados FM. A review of multiple-choice item-writing guidelines for classroom assessment. *Appl Meas Educ* 2002; 15:309–333
27. van der Gijp A, Ravesloot CJ, van der Schaaf MF, et al. Volumetric and two-dimensional image interpretation show different cognitive processes in learners. *Acad Radiol* 2015; 22:632–639
28. Lewis PJ, Chen JY, Lin DJ, McNulty NJ. Radiology ExamWeb: development and implementation of a national web-based examination system for medical students in radiology. *Acad Radiol* 2013; 20:290–296
29. Krupinski EA, Becker GJ, Laszakovits D, Gerdeman AM, Evanoff MG. Evaluation of off-the-shelf displays for use in the American Board of Radiology maintenance of certification examination. *Radiology* 2009; 250:658–664
30. Ben-David MF. AMEE guide no. 18: standard setting in student assessment. *Med Teach* 2000; 22:120–130
31. Cronbach LJ. Coefficient of alpha and the internal structure of tests. *Psychometrika* 1951; 16:297–334
32. Tavakol M, Dennick R. Post-examination interpretation of objective test data: monitoring and improving the quality of high-stakes examinations: AMEE guide no. 66. *Med Teach* 2012; 34:e161–e175

FOR YOUR INFORMATION

This article is available for CME and Self-Assessment (SA-CME) credit that satisfies Part II requirements for maintenance of certification (MOC). To access the examination for this article, follow the prompts associated with the online version of the article.