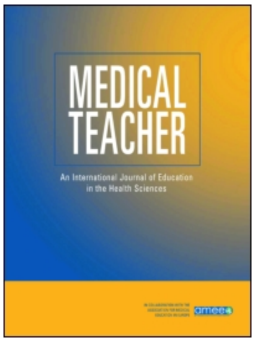
MEDICAL
TEACHER
An International Journal of Education in the Health Sciences

Taylor & Francis
Taylor & Francis Group

Medical Teacher

ISSN: 0142-159X (Print) 1466-187X (Online) Journal homepage: http://www.tandfonline.com/loi/imte20

The foundations of measurement and assessment in medical education

Mohsen Tavakol & Reg Dennick

To cite this article: Mohsen Tavakol & Reg Dennick (2017): The foundations of measurement and assessment in medical education, Medical Teacher, DOI: 10.1080/0142159X.2017.1359521

To link to this article: http://dx.doi.org/10.1080/0142159X.2017.1359521

Published online: 02 Aug 2017.

MEDICAL TEACHER, 2017
https://doi.org/10.1080/0142159X.2017.1359521

AMEE GUIDE

The foundations of measurement and assessment in medical education*

Mohsen Tavakol[a] and Reg Dennick[b]

[a]Medical Education Unit, Educational Development Center, The University of Nottingham, Nottingham, UK; [b]Medical Education Unit, The Medical School, The University of Nottingham, Nottingham, UK

ABSTRACT
As a medical educator, you may be directly or indirectly involved in the quality of assessments. Measurement has a substantial role in developing the quality of assessment questions and student learning. The information provided by psychometric data can improve pedagogical issues in medical education. Through measurement we are able to assess the learning experiences of students. Standard setting plays an important role in assessing the performance quality of students as doctors in the future. Presentation of performance data for standard setters may contribute towards developing a credible and defensible pass mark. Validity and reliability of test scores are the most important factors for developing quality assessment questions. Analysis of the answers to individual questions provides useful feedback for assessment leads to improve the quality of each question, and hence make students' marks fair in terms of diversity and ethnicity. Item Characteristic Curves (ICC) can send signals to assessment leads to improve the quality of individual questions.

Introduction

The DNA of any formal education is assessment. It is a systematic process that collects and interprets information derived from exam data to legitimize examination content and student marks. To achieve this assessment, leads need to provide evidence of the quality of assessment instruments. Psychometric methods allow us to detect poorly developed and performing assessment questions in order to improve their quality (Tavakol and Dennick 2016). Psychometric results enable assessment developers not only to develop fairer assessment questions in terms of equality and diversity, but also to improve the effectiveness of their teaching approaches. In addition, the psychometric analysis of assessment questions will enable medical educators to improve their skills in test development (Ebel 1972).

It should be emphasized that assessment is the *measurement* of learning and that an understanding of the factors that influence the accuracy, reliability, and validity of the measurement process are essential for the creation of high quality assessments. The purpose of this Guide is to provide a general introduction to the foundations of measurement and assessment in medical education for those new to the subject. It will cover the following topics: measurement and assessment, formative and summative assessment, norm-referenced and criterion-referenced assessments, standard setting, evidence of reliability and validity and choosing the best assessment questions.

Practice points

- Medical educators should become familiar with the foundations of measurement and assessment.
- Medical educators measure student learning and ability by means of formative and summative assessments to improve the quality of student learning and to improve the quality of the curriculum, teaching, and assessment.
- Criterion-referenced measurement is concerned with achieving the learning objectives of the curriculum.
- Standard setters should provide a reliable pass mark in order to divide students into two groups: competent and incompetent.
- Providing standard setters with feedback about their ratings may reduce the error attached to the pass mark.
- Both compensatory and conjunctive scoring may be used for moderating student marks.
- Reliability and validity are monitored to ensure the appropriateness and accuracy of student marks.
- Item characteristic curves illuminate the expected item mark as a function of student ability.
- Option characteristic curves illuminate the quality of effective options in multiple choice questions.

Measurement and assessment

Measurement

It has been defined as the assignment of numbers to objects, events, attributes, and traits according to rules (Miller et al. 2013). In this definition these characteristics are labeled by numbers. An example may clarify the "rules". If students take an exam with the same instructions, administration, assessment questions, and scoring system, we can compare students' marks with each other. For example, if a student receives a mark of 70% in physiology and a mark

CONTACT Mohsen Tavakol mohsen.tavakol@nottingham.ac.uk The University of Nottingham – Medical Education Unit, School of Medicine, Nottingham, UK

*This paper forms part of AMEE Guide No. 119 which will be available from www.amee.org of 80% in anatomy we have meaningfully measured the ability of the student. By assigning numbers to students, we have measured their performance and, marks signify differences in the characteristic being measured.

Assessment

Assessment is concerned with "How well does the individual perform?" (Miller et al. 2013). We measure to improve all assessment questions in order to ensure the accuracy and stability of the results. The assessment data are used for quality assuring the pass/fail decision of a cohort of students, the effectiveness of a course and the validity and reliability of the tests. Such pass/fail decisions are based on the measurement process. Valid and reliable assessments that measure the ability of students have three main goals: "to optimize the capabilities of all learners and practitioners by providing motivation and direction for future learning; to protect the public by identifying incompetent physicians; and to provide a basis for choosing applicants or advanced training" (Epstein 2007). In medical education, it is noteworthy to mention that assessment should be based on the learning outcomes of the individual courses which are themselves subject to national standards. In addition to these, post examination analysis of exam data can monitor and improve the exam cycle as described in Tavakol and Dennick, AMEE Guide 54.

Formative and summative assessment

Formative assessment

Students should be aware of their competency gaps or educational needs, the difference between their current status and their desired goals, and they should take action in order to achieve this (Black and Wiliam 1998). Formative assessment, sometimes called "assessment for learning", is an ongoing process or guide, not a formal test and is aimed at monitoring learning during teaching. The quality of teaching and the learning experiences are judged by formative assessment. Based on these judgments, medical teachers adjust educational materials and clarify learning outcomes in order for students to achieve the desired learning goals. Constructive feedback to students and educators is the cornerstone of formative assessment (Shepard 2006).

Summative assessment

Unlike formative assessment, the aim of summative assessment is to make an accurate pass/fail decision about students. Post-examination analysis of exam data can provide feedback for medical educators to gain understanding of the fitness of learning outcomes and the productivity of teaching. In addition, by summative assessment, we assure the public that our students have minimum standards for the diagnosis and treatment of patients (Norcini and Dawson-Saunders 1994).

Norm-referenced and criterion-referenced measurement

These two terms are widely used in medical education, and both are concerned with the interpretation and decision of the assessment results. *Norm-referenced interpretations* are concerned with a student's mark relative to the distribution of marks of a cohort of students. The performance of individual students is compared with the cohort using an arbitrary number as the pass mark or grade boundary. Those who have a mark equal or greater than the pass mark will pass the exam but this may not reflect the acquisition of the learning outcomes of the curriculum. Whether a student passes or fails, or achieves a particular grade, depends on the performance of the cohort and not on the individual. However, norm-referenced assessments can contain hard questions in order to differentiate high and low performers. This is useful for selecting applicants when there are limited positions available, such as selection for jobs or admission to further study.

Criterion-referenced interpretations, sometimes called objective referenced, are concerned with the criteria forming the learning outcomes of a course. In this approach, a student's mark is interpreted based on the achievement of learning outcomes without any comparison with other students. For example, if an exam has 20 questions and each question measures a specific learning outcome, if Rita answers 16 questions correctly, she will achieve 80% of the learning outcomes.

Standard setting

Setting performance standards is an important issue in medical education. Standard setters wish to know about students' abilities, and whether they are able to perform specific tasks in the real world. The aim of standard setters is to categorize students into appropriate groups such as pass, fail or borderline. Most standard setting methods use the estimated performance of a borderline student who is on the border between pass and fail to identify a pass mark that establishes the minimum level of performance, thereby discriminating between those students who fail from those who pass (Kolen 2006).

A number of standard setting methods have been described. For a more complete description of these methods, see Cizek (1996), Downing et al. (2006), and McKinley and Norcini (2014). Standard setters estimate the probability that a borderline student will answer the items correctly. The most popular test-centered methods are the Angoff method (and its modifications), the Ebel method, and the Nedelsky. These methods have been criticized for two reasons. First, it is very difficult for standard setters to imagine the knowledge and skill levels of borderline students in order to estimate the probability that they answer an item correctly. Secondly, if standard setters are changed, the pass mark will change (Cizek 1993). In student-centered methods, the pass mark is based on students' actual performance on a specific assessment. In these methods, students' performances are scrutinized rather than the assessment questions. The most popular test-centered methods are the median borderline method, the regression model and the contrasting group method.

Presentation of performance (normative) data

Do we need to provide standard setters with performance data for each item prior to setting a pass mark? It seems that there is controversy surrounding the influence of performance data on the pass mark. A meta-analysis showed that presenting the item difficulty values to standard setters resulted in low pass mark using Angoff's methods (Hurtz and Auerbach 2003). It has been argued that providing item difficulty values impacts on the variability among standard setters rather than on the established pass mark. In addition, standard setters "feel more confident about the resulting performance standards if there has been discussion and feedback" (Hambleton et al. 2012). An experimental study suggested that standard setters match initial ratings with the performance data (Clauser et al. 2009). Some studies show that the pass mark increases or decreases by providing performance data to standard setters. Additionally, it has been documented that providing standard setters with feedback using different approaches on their ratings can enable them to identify their standard setting errors.

The compensatory and conjunctive standard-setting strategies

Compensatory strategy/scoring refers to the sum of a battery of assessments which are compared with a particular pass mark to make a pass/fail judgment. For example, in an OSCE examination, if we have 20 stations, with an overall pass mark of 65%, if the average of all stations' scores are calculated then those who have received a score of 65% or greater would pass the whole OSCE. In this approach those who have received a low score on some stations but have received high scores on other stations can compensate their low scores and may pass the OSCE. This strategy is useful for moderating student marks if the moderating committee (content experts) found that some stations had problems, for example, examiner unreliability. If stations all measure a single construct, such as the construct of clinical performance, the average of the station scores meaningfully represents the construct of interest, and hence a low score on one or two stations can be overlooked if overall performance is adequate (Haladyna and Hess 1999).

In conjunctive scoring, each station constitutes a single construct with a separate pass mark, and failing these stations is not tolerated since each station is necessary for patient safety. Scores on one station do not influence whether a student passes other stations. For professional certification and licensure tests, assessment leads can use conjunctive scoring as they believe that a licentiate should be competent in the construct of interest. Consequently, the sum of the stations scores does not make sense in conjunctive scoring. Clearly, fails will be greater in conjunctive scoring than in compensatory scoring. Although the conjunctive approach is central to the legitimation of a physicians' competency and capability, this strategy will potentially result in more failures, which might be professionally problematic (Haladyna and Hess 1999). However, it should be emphasized that the compensatory approach may make better sense in many subject areas as students may not be equally good enough at every competency, and hence strengths can compensate for weaknesses (Zieky and Perie 2006).

Reliability and validity

Reliability and validity are two important aspects of measurement. An assessment can yield a reliable score if and only if a cohort of students can be consistently rank-ordered when the assessment is administered on different occasions. A useful analogy for understanding reliability is that of "noise" in a test. Anything that detracts from the measurement taking place will create error and noise in the test and consequently will add to unreliability. There are different approaches for measuring the reliability of test scores. Some of these approaches are test-re-test reliability, parallel form, split-half, coefficient alpha, and Kuder–Richardson, Hoyt's method (which is estimated using the analysis variance approach), Coefficient theta (using factor analysis), Omega, Inter-rater reliability (agreement), and Generalizability theory. It is noteworthy to mention that if an assessment does not have an acceptable reliability, there will always be a question mark over its utility. More importantly, an assessment may consistently rank-order a cohort of students, but this does not say anything about what it intends to measure. Consequently, validity is another property of a test which must be considered.

Validity

Validity is concerned with "the degree to which evidence and theory support the interpretation of test scores entitled by proposed uses of tests" (American Educational Research Association (AERA) 1999). Given this, assessors should be clear about the proposed interpretation and use of student marks. For example, if we think a student has achieved a distinction by correctly answering 95% of the learning objectives in a test, have they really done this? Or has someone directly helped the student during the assessment? If some questions were not based on the learning objectives or if some questions were not differentiated between high and low performers, the 95% may not be a good indicator for interpreting student performance. It should be emphasized that validity is neither concerned with assessment questions nor the assessment results. It is concerned with the inferences and decisions of the assessment results (Kane 2002).

Traditional validity types which included content-related validity, criterion-related validity, and construct-related validity have now been discarded by the *Standards for educational and psychological testing* (American Educational Research Association (AERA) 1999), referred to as *Standards*. Instead, five types of validity evidence have been described by this document, which are discussed as follows.

Evidence based on assessment content

Assessment questions are a sample of all potential assessment questions and hence we need to investigate how well the sample of assessment questions can be generalized to all possible assessment questions. How well do the assessment questions align with the learning objectives? How well do the assessment questions represent the domain of interest? How well are the assessment questions formatted, written and thematized. How well are the assessments administered and scored? Those who have expertise in the content domain can provide evidence based on content. For example, a panel of experts can interpret the representativeness of a sample of assessment questions on a test for a given cohort of students.

Evidence based on response process

This type of validity requires evidence on how much the construct being measured fits the nature of performance or response in which students are engaged. For example, if assessors developed a test to measure the construct of depression, do the assessment questions fit the construct of depression, i.e. construct representation. Do assessment questions associate with other factors which are not concerned with the construct of depression, i.e. construct-irrelevant variance? Therefore, the validity of a test will be threatened, if the construct of interest is underrepresented or influenced by irrelevant factors. Interested readers can find further information about construct underrepresentation and construct-irrelevant variance elsewhere (Downing 2002).

There are different methods for obtaining validity evidence for the response process. These methods are based on qualitative data collection methods such as think aloud interview and focus group interviews. For example, observing examiners in OSCEs allows us to understand how they rate and interpret the performance of students. Assessors should ensure that the examiners rate students based on the intended criteria rather than irrelevant factors.

Evidence based on internal structure

This type of evidence validity is concerned with quantitative methods using psychometric-statistical inferences. We want to provide evidence of the association between items and assessment results and the construct being measured. An assessment may measure a single construct (unidimensional) or multiple constructs (multidimensional).

An assortment of methods can be applied to establish evidence based on internal structure. For example, Rasch analysis is one method that enables us to identify the psychometric structure of assessment questions. Using factor analysis, we can identify the internal structure of assessments. Another approach is called the contrasted groups approach, sometimes also called the known-group approach. Here, the test is administered to two groups of people who have different knowledge of the construct of interest (extremely high and extremely low).

Evidenced based on relations to external variables

Another approach suggested by the AERA for providing validity evidence is to identify the association between test scores with external variables. The scores of two assessments are correlated with each other if both measure the same construct. We would expect to obtain a positive correlation between a communication skills test and a psychiatry test, and perhaps a low correlation between a communication skills test and a surgery test. In addition, we may also wish to predict the performance of the students in the future. For example, in the admission process, if scores on performance in physics and mathematics are highly correlated with the later performance in medicine, the admission leaders may consider physics and mathematics as entry requirements for medicine. Here, physics and mathematics are called criteria and this approach is called the criterion-related validity. There are two types: concurrent and predictive. *In concurrent validity*, the test and the criterion are administered simultaneously. The correlation between the test scores and the criteria are calculated as reflecting a validity coefficient. For example, we could administer a situational judgment test (SJT) and a particular OSCE station during the same exam day and then correlate these scores. Predictive validity involves using the test scores to predict the behavior of students in the future. To obtain the predictive validity coefficient the test is administered and after collecting the criterion scores (e.g. after 6 months), we are in a position to calculate the correlation between the test scores and the criterion scores. For example, the lead admissions tutor of a medical school may consider physics A-level as a good predictor of medical school performance (the criterion), if a good correlation is found.

Evidence based on consequence of testing

The last type of validity evidence explained in the *Standards* is based on the intended and unintended outcomes of assessment results. How can assessment questions influence the education system as whole? The intention of both formative and summative assessments is to improve student ability. But do they improve the ability of students? Do they enhance student motivation? Do they improve teaching? Answering these questions can provide positive consequence validity evidence of assessment results. The interpretations of assessment results may have a number of unintended negative outcomes, e.g. increasing the dropout rates of medical students or focusing on the test while ignoring learning objectives. Equally, consider the development of a test to measure student performance on gynecology and obstetrics. If the test is biased and female students outperformed male students, an unintended consequence may occur.

The analysis of assessment questions

The analysis of assessment items provides useful information about the marks that students have received from their exams. Student marks can be misleading if errors are attached to them. Poor quality assessment items can be a source of error in generating an unfair mark. If items are incorrectly assigned the wrong answer by the test writer; if items have more than one best answer or if items are too hard for a cohort of students we will get a misleading mark. These items, sometimes called underperforming/rogue/flawed items, should be adjusted before students' marks are published. Two common statistics are usually used to identify the underperforming items: item difficulty and the item discrimination index. Item difficulty refers to the proportion of students who get the question right. Item discrimination indicates whether or not the items differentiate high and low performers. Consider if 30% of students answered item1 correctly, and 70% of students answered item 2 correctly, then item 2 was easier than item 1. If the value of the item difficulty is close to 0 or 1, the item needs to be examined and perhaps discarded as it

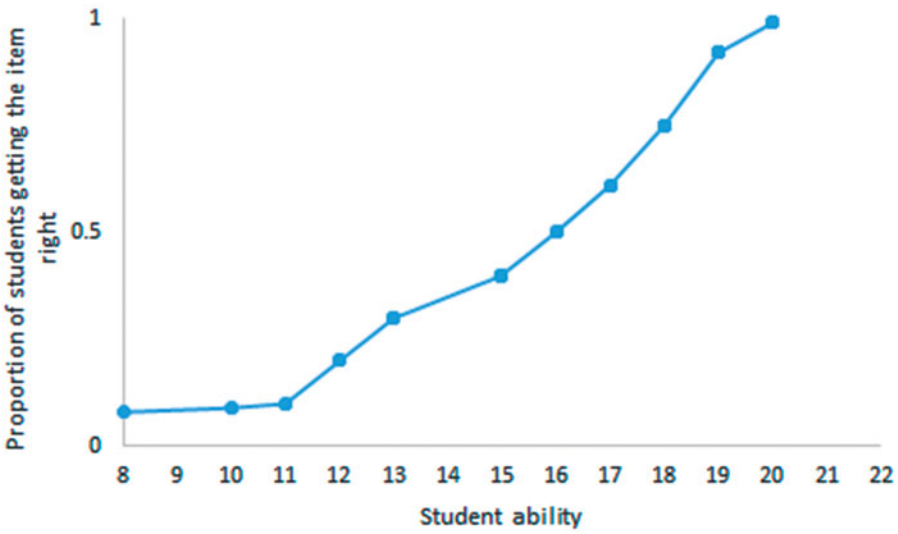

Figure 1. Item characteristic curve of an estimated item.

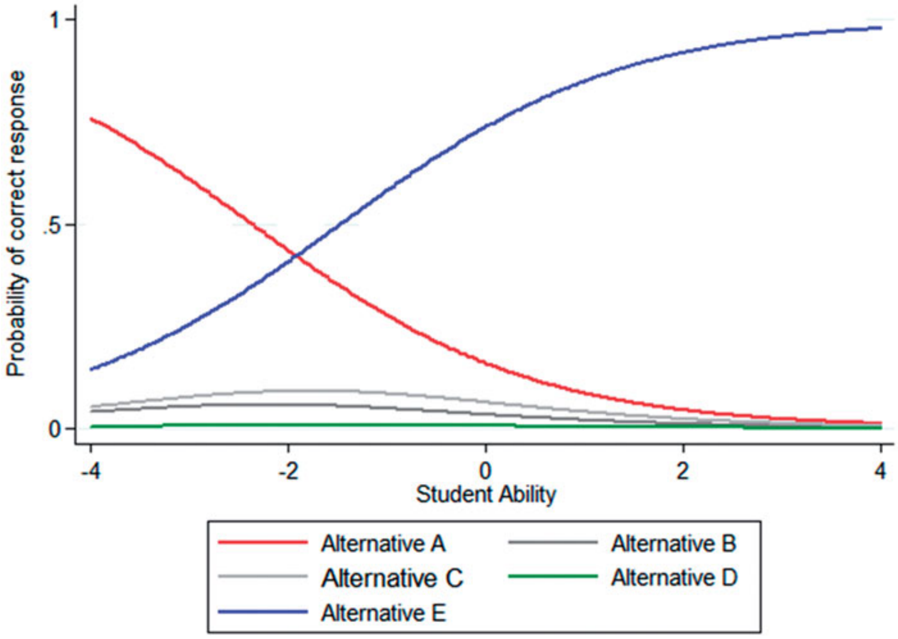

Figure 2. Trace lines for five alternatives. Alternative E is correct.

does not provide any information about the differences between the ability of students. Too easy and too difficult items do not differentiate students in terms of the performance being measured.

How should we judge the quality of assessment items? From a psychometric perspective, an item has good quality if it has a high item discrimination index. A variety of approaches are used to calculate the item-discrimination index. It has been well documented that the point-biserial correlation (the correlation between item score and the total mark excluding the item score) provides the best indicator of the item quality (Kelley et al. 2002). A good item has a point-biserial correlation of 0.25 or above. A negative value of the discrimination index indicates those who performed poorly on assessment answered the item correctly. Such items may indicate that an error (noise) occurred in the student marks. For example, the item may be miskeyed or poorly constructed. Such items should be revised or discarded.

Item characteristic curves (ICC)

ICC illustrate the relationship between student ability and item difficulty (the proportion of students answering an item correctly) of a test. To plot an item characteristic curve, students' marks are placed on the horizontal axis and the value of item difficulty on the vertical axis (see Figure 1). As you can see from this figure, this item has discriminated students soundly meaning that those who performed poorly on the whole test answered the item incorrectly.

Option characteristic curve

The analysis of students' responses to correct and incorrect options in assessment questions provides useful information about the plausibility of options and the effectiveness of questions (Schmeiser and Welch 2006). The frequency distribution of students' responses about the option responses in a question can be analyzed to judge the plausibility of options in multiple choice assessments. The effectiveness of options can be assessed using the point-biserial correlation, that is, correlating the option responses with total mark. If there is a negative correlation between the correct option and the total test mark, then the item has a fundamental problem as low performers get the question right, but high performers get the question wrong. A functional distractor (plausibly incorrect item) has a negative correlation with the total test score, if it is indeed a distractor. If a wrong option is not chosen by students (high and low performers), the option should be excluded from the question. A functional distractor should have a distribution frequency of greater than 5% for a cohort of students (Haladyna and Downing 1988).

Option characteristic curves or trace lines can portray functional and dysfunctional distractors in a question. Figure 2 shows that the trace lines in a multiple-choice question from a cohort of students. Alternative A shows that the tendency towards the selection of this alternative was decreased as student ability was increased. Alternatives B, C and D were selected by few students reflecting that these three incorrect alternatives were not plausible and easily eliminated meaning that this item did not discriminate between high and low performers. Correct alternative E was selected by the majority of students and selecting this alternative became high as student ability increased.

Conclusions

This AMEE Guide indicates the importance of measurement and assessment in teaching and learning in medicine. Measurement deals with “How much” while assessment is concerned with the measurement of student learning and ability. We measure this learning in order to improve the quality of the curriculum. On one hand, formative assessments contribute towards teaching, learning, and feedback in order to improve the learning outcomes of a specific course. Summative assessments, on the other hand, contribute towards the legitimation of student performance. This requires that standard setters provide a credible and defensible pass mark in order to split students into two groups: masters and non-masters. Providing standard setters with feedback using different standard setting methods on their ratings will minimize the errors attached to their ratings.

Assessment leads should provide evidence of reliability and validity in order to accurately interpret student marks which in turn leads to an improvement in the assessment of learning objectives in a specific test. Using different approaches, especially Omega, assessment leads are able to provide evidence of reliability for their assessments. Option characteristic curves show the frequency distribution of students’ responses to the alternative options in a question.

Disclosure statement

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this article.

Notes on contributors

Mohsen Tavakol, PhD, MClinEd, is assistant professor of psychometrics in the Medical School at the University of Nottingham. His main interests are in medical education assessment, psychometric analysis (Classical Test Theory, Generalisability theory, Item Response Theories Models, standard setting, roust statistical methods, multivariate statistics, multilevel modelling, quantitative and qualitative research methods.

Reg Dennick, PhD, MEd, FHEA, is a professor of medical education in the University of Nottingham. His main interests are in medical education teaching and research, problem-based learning, assessment, clinical reasoning, staff training and curriculum development.

References

American Educational Research Association [AERA]. 1999. American Psychological Association & National Council on Measurement in Education. *The standards for educational and psychological testing*, Washington (DC): American Educational Research Association.

Black P, Wiliam D. 1998. Assessment and classroom learning. Assess Educ: Principles Policy Pract. 5:7–73.

Cizek G. 1993. Reconsidering standards and criteria. J Educ Meas. 30:93–106.

Cizek G. 1996. Setting passing scores. Educ Meas: Issues Pract. 15:20–31.

Clauser B, Mee J, Baldwin S, Margolis M, Dillon G. 2009. Judges' use of examinee performance data in an Angoff standard-setting exercise for a medical licensing examination: an experimental study. J Educ Meas. 46:390–407.

Downing S. 2002. Threats to the validity of locally developed multiple-choice tests in medical education: construct-irrelevant variance and construct underrepresentation. Adv Health Sci Educ. 7:235–241.

Downing S, Tekian A, Yudkowsky R. 2006. Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. Teach Learn Med. 18:50–57.

Ebel R. 1972. Essentials of educational measurement. London: Prentice-Hall International, Inc.

Epstein R. 2007. Assessment in medical education. N Engl J Med. 356:387–396.

Haladyna T, Hess R. 1999. An evaluation of conjunctive and compensatory standard-setting strategies for test decisions. Educ Assess. 6:129–153.

Haladyna TM, Downing S. 1988. *Functional Distractors: Implications for Test-Item Writing and Test Design*. [accessed 2015 Aug 10]. http://files.eric.ed.gov/fulltext/ED293851.pdf.

Hambleton R, Itoniak M, Copella J. 2012. Essential steps in setting performance standards on educational tests and strategies for assessing the reliability of results. In: Cizek G, editor. Setting performance standards. London: Routledge.

Hurtz G, Auerbach M. 2003. A meta-analysis of the effects of modification to the Angoff method on cut-off scores and judgment consensus. Educ Psychol Meas. 63:584–601.

Kane M. 2002. Validating high-stakes testing programs. Educ Meas: Issues Pract. 21:31–41.

Kelley T, Ebel R, Linacre J. 2002. Item discrimination indices. Rasch Meas Trans. 16:883–884.

Kolen M. 2006. Scaling and norming. In: Brennan R, editor. Educational measurement. Westport (CT): American Council on Education.

Mckinley D, Norcini J. 2014. How to set standards on performance-based examinations: AMEE Guide No. 85. Med Teach. 36:97–110.

Miller M, Linn R, Gronlund N. 2013. Measurement and assessment in teaching. Boston: Pearson.

Norcini J, Dawson-Saunders B. 1994. Issues in recertification in North America. In: Newble D, Jolly B, Wakeford R, editors. The certification and recertification of doctors. Cambridge: Cambridge University Press.

Schmeiser C, Welch C. 2006. Test development. In: Brennan RL, editor. Educational measurement. USA: American Council on Education.

Shepard L. 2006. Classroom assessment. In: Brennan R, editor. Educational measurement. Westport (CT): American Council on Education.

Tavakol M, Dennick R. 2016. Post-examination analysis: a means of improving the exam cycle. Acad Med. 91:1324.

Zieky M, Perie M. 2006. *A primer on setting cut scores on tests of educational achievement*. ETS. [accessed 2016 Jun 10]. https://www.ets.org/Media/Research/pdf/Cut_Scores_Primer.pdf.