PERSPECTIVE

Assessment in Medical Education: A Primer on Methodology

By **Anna Reinert, MD**
College of Physicians and Surgeons, Class of 2013

Good assessment is a major challenge within medical education. Among other factors, one obstacle to sound methods of assessment may be a lack of familiarity on the part of medical educators with the concepts of psychometrics that underlie assessment [1]. This primer gives an overview of the basic ideas and vocabulary that one should understand in order to evaluate the quality of any assessment tool designed for the purpose of evaluating medical students or residents. A popular conceptual framework for categorizing methods of assessment in medical education is described, and methods currently in use within each category are reviewed with attention to their psychometric strengths and weaknesses.

Assessment in medical education serves multiple purposes. It can be *formative*: promoting reflection, guiding future learning and shaping values – or *summative*: judging an individual's cognitive achievement or clinical performance [2]. Formative assessment can be employed as a tool for students to identify and respond to their own learning needs; to this end, results should profile distinct areas of strength and weakness [2, 3]. Summative assessments typically give results as a single aggregated score, and may be used to competitively rank trainees[3, 4], or to ensure quality control of trainees by screening for a minimal standard of competency – this latter goal of assessment being crucial to the maintenance of public trust in the profession of medicine [2, 5]. Many summative assessments serve both purposes. It is widely recognized that assessment drives learning: medical students study in order to pass (and excel on) exams [6-9]. Hence, an assessment should be designed to direct study in the most relevant way [5]. Some believe that the best way to achieve this relevance is by employing assessment formats that approximate professional reality [7]: for instance, by simulating the tasks and decisions faced by a doctor in the course of their work.

A *test construct* is "the concept or characteristic that a test is designed to measure" [8]. Within contemporary medical education, a set of constructs commonly employed is the Accreditation Council of Graduate Medical Education (ACGME) Core Competencies. Although the ACGME competencies pertain to the training of residents and fellows, they have been widely adapted as education standards for use at the medical school level as well. Six general competencies were advanced in 1999: *Patient care, Medical knowledge, Practice-based learning and improvement, Interpersonal and communication skills, Professionalism, and Systems-based practice*. The six core competencies are described in Table 1.

Devised to be applicable to all physicians, the ACGME core competencies were put forth with the goal of making medical education more outcome-driven [11]. The competencies are intended to represent habits rather than accomplishments - their assessment aims to "provide insight into actual performance" [2]. Achieving this relevance through the use of standardized assessment instruments remains a major challenge within medical education [11] and is limited by the fact that the majority of existing standardized assessment instruments focus on "cognitive achievement" rather than on performance or behavior [4].

To better understand the distinction between cognitive and behavioral methods of assessment, it is helpful to reference Miller's Pyramid. Introduced into the medical education literature in 1990, this framework differentiates four categories of assessment methods on the basis of what is required of the trainee (Figure 1). The pyramid was annotated by Crossley *et al*, who noted that the categories of *Knows* and *Knows How* are measures of cognition, whereas *Shows How* and *Does* imply assessment of behavior. Successive levels of the pyramid build upon lower levels: one cannot assess clinical competence without simultaneously assessing knowledge [6].

Higher levels of the pyramid are considered to more closely approximate professional reality but should not be considered superior, because "superior methods are those best aligned with the purpose of the test"

Table 1. ACGME Core Competencies [10]	
Patient Care	The ability to provide patient care that is compassionate, appropriate, and effective for the treatment of health problems and the promotion of health.
Medical knowledge	Demonstration of knowledge of the established and evolving biomedical, clinical, epidemiological and social-behavioral sciences, as well as the application of this knowledge to patient care.
Practice-based learning and improvement	The ability to investigate and evaluate one's care of patients, to appraise and assimilate scientific evidence, and to continuously improve patient care based on constant self-evaluation and life-long learning.
Interpersonal and communication skills	Demonstrates interpersonal and communication skills that result in the effective exchange of information and collaboration with patients, their families, and health professionals.
Professionalism	Demonstration of a commitment to carrying out professional responsibilities and an adherence to ethical principles.
Systems-based practice	Demonstrates awareness of and responsiveness to the larger context and system of health care, as well as the ability to call effectively on other resources in the system to provide optimal health care.

[5]. The challenge of performance-based assessment (*Shows, Does*) is that it is more difficult to design and administer with the degree of objectivity, validity and reliability that is considered essential to fair and effective assessment [12]. For this reason, cognitive achievement-based methods of assessment (Knows, Knows How) are often better suited to the ranking of examinees, whereas performance-based methods of assessment (Shows, Does) are more psychometrically limited and may only have validity for use in screening for a minimal level of competency or in identifying outstanding trainee performance [12].

Validity refers to the meaningfulness of the interpretation of an assessment result [8, 13]; it is considered necessary to the sound, ethical, and effective use of tests [8]. The use of invalid tests can lead to decisions that are ineffective or unfair [14]. Reliability is a necessary component of validity [3]: it refers to the consistency and precision of test measurements, also described as score reproducibility [15, 16]. Error in measurement reduces reliability and "reduces the confidence that can be placed in any single measurement" [8], as one cannot be confident that a second administration of the same assessment would result in that same score.

There are two common, general threats to validity: construct-irrelevant variance and construct underrepresentation [13]. Poorly written test items, cheating, and differences in student knowledge about test content result in construct-irrelevant variance, defined as the "degree to which test scores are affected by processes that are extraneous to its intended construct" [8]. This variance produces systematic (non-random) error in score measurement, and distorts the meaning of test scores [8, 13]. Appropriate test security and elimination of biased test items will reduce construct-irrelevant variance and improve validity: these priorities require vigilance and continual quality assurance and improvement by an educational program for the purpose of maintaining exam validity.

In the same way that higher levels of Miller's Pyramid build upon lower levels, the demonstration of competence is contingent upon knowledge of the clinical content in which the assessment is framed. A student may demonstrate excellent problem-solving ability on one exam question but fail to exhibit the same ability in responding to a different question with content about which they are less knowledgeable. This phenomenon, known as case specificity [7], results in construct underrepresentation. Defined as the "degree to which a test fails to capture important aspects of the construct" [8], construct underrepresentation is a limitation to validity that especially affects assessments based on a limited number of cases. Too-few test items may result in inadequate sampling of content from certain domains of knowledge [13, 17]. Use of faculty "pet topics" on an exam may also result in construct underrepresentation – especially if faculty members also "teach to the test" [18]. This threat to validity can be minimized through careful exam design.

When alternate forms of an exam are used, the chance that certain examinees will tend to score better on one form of an exam than on another constitutes a threat to reliability. This common obstacle to reliability is known as examinee-by-task interaction [7, 17]. Akin to case specificity, it can be minimized by increasing the number of cases on an assessment [3, 19]. Another threat to reliability is subjectivity in grading, which limits what is known as inter-rater reliability and is a threat more broadly to the objectivity of the assessment. Defining clear guidelines for exam grading through a detailed grading rubric is the best way to achieve inter-rater reliability.

Issues of validity and reliability are a major differentiating factor between methods from the four categories of Miller's Pyramid. To illustrate these differences, we review four varieties of assessment tools common within contemporary medical education, each representing a different level of the Miller's Pyramid: the *Multiple Choice Question* examination to represent *Knows* methods, the *written/computer simulation* examination to represent *Knows-How*, the *OSCE* to represent *Shows*, and *Chart Review* to represent *Does*.

Multiple Choice Question (MCQ) examinations assess what a medical trainee *Knows* [5]. They have been historically favored in medical education due to their efficient sampling of many clini-

cal situations and content areas and their ease of administration and grading [2]. MCQ exams may be efficiently administered to a large number of examinees and have been shown to have excellent reliability [7], making it a fair and effective tool for stratifying individual student performance. Critics deride this exam format for being poorly linked to professional reality: testing recall of isolated facts that constitute trivial knowledge [7, 20]. Modern exam developers have taken note of these critiques and most MCQs are now framed within clinical scenarios; yet, while "a well-constructed MCQ may demand a great deal of analytical thinking," [6] the response format does not permit assessment of a trainee's deliberation and reflection [21]. Furthermore, the response format of a MCQ cannot differentiate between correct answers based on knowledge and those based on random guessing. Because identifying the correct answer to an MCQ requires only recognition of the best answer among an array of choices, these exams test "recognition memory" rather than "free recall" [21, 22]. Termed the *cueing effect*, this phenomenon has been shown to overestimate student knowledge [15, 22]. The consequence of cueing is that students, being required only to identify correct answers rather than to explain how they arrived at a response, may be driven to rote learning rather than understanding [6]. Rote learning does not prepare a student as well as true understanding does for the responsibilities of being a physician.

Written/Computer simulation examinations assess whether a medical trainee *Knows How* to approach a clinical challenge [5, 12]. Several exam formats emerged in the 1960s as educators strove to measure students' clinical reasoning, problem solving, and clinical judgment [7, 9]. The *Patient Management Problem (PMP)* was the first popular exam of this variety; it required that a student work through a clinical problem by making diagnostic and management decisions, often taking the form of a branched pathway through the problem with outcomes and successive answer choices that depended on the previous choices being made [6, 7]. Though briefly adopted for the purpose of licensing and specialty examinations, issues of content-specificity resulting in low reliability led to cessation of use of the *PMP* by the NBME and American Board of Internal Medicine [21]. Content-specificity was a major limitation for the *PMP* because each clinical case contained several questions and was time-consuming to complete, with the consequence that only a few cases were sampled for each exam administration.

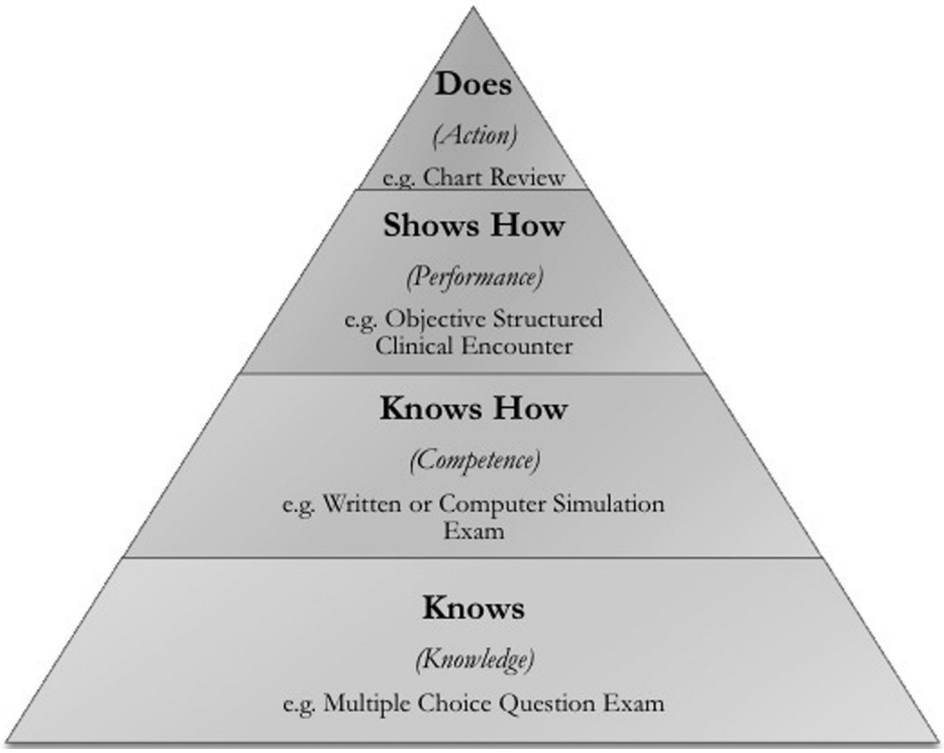

Figure 1. Miller's Pyramid Framework for Clinical Assessment. Adapted from Miller GE. The Assessment of clinical skills/competence/performance.

The *Key Features* exam introduced an innovation that allowed for sampling of a larger number of clinical topics within a single exam administration by testing only the "essential elements in decision making", "the critical steps in the successful resolution of the clinical problem" [23]. Developed in 1987, this exam format features write-in and short menu formats, and asks only a limited number of questions about a case, all of them directed at essential "key features" of the case [23, 24]. Thanks to shorter problem length and more clinical scenarios per exam, the *Key Features* exam format has less content-specificity and improved reliability over the *PMP* [6]. However, given the use of short menu (multiple choice question) items, the *Key Features* exam - like a MCQ test - is subject to the cueing effect.

Written simulation examinations such as the *PMP* and *Key Feature exam* serve to measure clinical judgment or reasoning and problem solving abilities of examinees. Professional realism of these exams is good, as

questions resemble actual decisions faced by practitioners in the course of patient care [20]. Compared to MCQ exams, these exams are less influenced by cueing and do not overestimate examinee ability [15]. The major limitation of these examinations is case-specificity, which results in lower reliability than may be achieved with an MCQ exam [6,7]. The written simulation exam format holds great potential, and innovation in methodology may yet yield an exam format with improved reliability. Inevitably, however, written examinations may not be able to assess "the technical and performance aspects of physical diagnosis, communication, humanism and professionalism" [18]. To evaluate performance and execution, we look to the methods from the upper two levels of Miller's pyramid.

The *Objective Structure Clinical Encounter (OSCE)* assesses whether a trainee can Show How they respond to a clinical challenge [5, 12]. Pioneered in the late 1970s, the exam format employs standardized patients and consists of multiple exam stations through which students rotate, each focusing on a single skill [7, 25]. Considered by Miller to be "the most effective substitute for reality" [12], this performance-based exam format is viewed by students as "more relevant to their subsequent work as interns" [26], and is considered best suited to the evaluation of technical skills [9]. Psychometrically, outcomes for the *OSCE* parallel those of written simulation exams, with content-specificity constituting a major limitation to reliability [7]. In addition, the *OSCE* has issues of non-consensus in scoring, resulting in poor inter-rater reliability [26]. Relative to written exams, the performance-based *OSCE* is more time consuming and resource intensive [2, 12]: factors which affect the practicability of employing this exam format within an education program.

Chart review is a method of assessing what a trainee *does* by reviewing their routine behavior [5]. Use of this method is rare and has not been subjected to the same rigorous psychometric studies as have the assessment methods discussed above [2]. When used, it often consists of peer review and judgment of whether the care provided by an individual meets the standard of "generally accepted practice". This criterion is useful for "identifying extreme individual deficiencies and widespread deviation" [4], but is largely subjective and has little value in the objective assessment of trainees for the purpose of stratifying performance.

In conclusion, each method (and each level of Miller's pyramid) plays an important role in obtaining an accurate view of the overall qualities and achievements of an individual trainee. Individual competencies may be best measured through distinct instruments, so it makes sense to use a combination of assessment methods [2, 4]. Conversely, every method of assessment has its own limitations, and "no single test is the panacea to the issue of assessment" [7]. Since "the use of multiple methods of assessment can overcome many of the limitations of individual assessment format" [2], and can serve to "broaden the type of evidence brought to bear on the determination of competence" [21], it should be considered preferable to reliance on a single method of assessment.

Miller's choice of a pyramid shape for framing his discussion of methodology is a statement about the ideal composition of a testing program, given practical and methodological realities: primary reliance on *Knows* and *Knows How* methods, with some smaller component of performance-based *Shows How* and *Does* methods of assessment. Citing Alfred North Whitehead's quip that "there is nothing more useless than a merely well informed man", Miller expresses a strong preference for innovation in methods at the *Knows How* level to balance the *Knows* level methods that dominate contemporary medical education [12]. His identification of the *Knows How* level of the pyramid as representative of "competence" suggests a natural relationship with the ACGME clinical competencies put forth in the decade following Miller's landmark paper. However, as Miller acknowledged, the assessment of "know-how" or competence is not equivalent to documentation of actual performance or habit [12]. Given that the ACGME competencies are intended to represent habits and to "provide insight into actual performance" [2], we can conclude that assessment of the ACGME competencies probably requires some combination of methods from the *Knows-How, Shows and Does* levels of Miller's pyramid. A priority within contemporary medical education should be the development of assessment tools that query student "know how", as this method is currently underrepresented with respect to its relative importance as prescribed by Miller.

Given the high stakes often ascribed to grades within medical school, the topic of assessment in medical education can provoke strong emotions. Hopefully this article serves to increase the transparency of assessment, and to highlight the challenges inherent in the process of fairly evaluating medical students. Perhaps the reader will even be compelled to become involved in the creation of new methods of objective assessment. It is an area of research ripe for innovation, and one that would benefit from increased involvement on the part of students.

References

1. Spencer, J., The metric of medical education. Medical Education, 2002. 36: p. 798-799.
2. RM, E., Assessment in Medical Education. New England Journal of Medicine, 2007. 356: p. 387-396.

(continued on 34)

2006.
3. Thompson, A. and Kent, G. Adjusting to disfigurement: processes involved in dealing with being visibly different. Clin Psychol Rev. 21: 663, 2001.
4. Murray, L., Arteche, A., Bingley, C., et al. The effect of cleft lip on socio-emotional functioning in school-aged children. J Child Psychol Psychiatry. 51: 94, 2010.
5. Sagheri, D., Ravens-Sieberer, U., Braumann, B., et al. An Evaluation of Health-Related Quality of Life (HRQoL) in a group of 4-7 year-old children with cleft lip and palate. J Orofac Orthop. 70: 274, 2009.
6. Kramer, F.J., Gruber, R., Fialka, F., et al. Quality of life and family functioning in children with nonsyndromic orofacial clefts at preschool ages. J Craniofac Surg. 19: 580, 2008.
7. Kramer, F.J., Gruber, R., Fialka, F., et al. Quality of life in school-age children with orofacial clefts and their families. J Craniofac Surg. 20: 2061, 2009.
8. Berk, N.W., Cooper, M.E., Liu, Y.E., et al. Social anxiety in Chinese adults with oral-facial clefts. Cleft Palate Craniofac J. 38: 126, 2001.
9. Wirls, C.J. and Plotkin, R.R. A comparison of children with cleft palate and their siblings on projective test personality factors. Cleft Palate J. 8: 399, 1971.
10. Brantley, H.T. and Clifford, E. Cognitive, self-concept, and body image measures of normal, cleft palate, and obese adolescents. Cleft Palate J. 16: 177, 1979.
11. Persson, M., Aniansson, G., Becker, M., et al. Self-concept and introversion in adolescents with cleft lip and palate. Scand J Plast Reconstr Surg Hand Surg. 36: 24, 2002.
12. Ramstad, T., Ottem, E. and Shaw, W.C. Psychosocial adjustment in Norwegian adults who had undergone standardised treatment of complete cleft lip and palate. II. Self-reported problems and concerns with appearance. Scand J Plast Reconstr Surg Hand Surg. 29: 329, 1995.
13. Starr, P. Self-esteem and behavioral functioning of teen-agers with oral-facial clefts. Rehabil Lit. 39: 233, 1978.
14. Clifford, E. Parental ratings of cleft palate infants. Cleft Palate J. 6: 235, 1969.
15. McWilliams, B.J. and Paradise, L.P. Educational, occupational, and marital status of cleft palate adults. Cleft Palate J. 10: 223, 1973.
16. Millard, T. and Richman, L.C. Different cleft conditions, facial appearance, and speech: relationship to psychological variables. Cleft Palate Craniofac J. 38: 68, 2001.
17. Neiman, G.S. and Savage, H.E. Development of infants and toddlers with clefts from birth to three years of age. Cleft Palate Craniofac J. 34: 218, 1997.
18. Noar, J.H. Questionnaire survey of attitudes and concerns of patients with cleft lip and palate and their parents. Cleft Palate Craniofac J. 28: 279, 1991.
19. Noar, J.H. A questionnaire survey of attitudes and concerns of three professional groups involved in the cleft palate team. Cleft Palate Craniofac J. 29: 92, 1992.
20. Richman, L.C. Parents and teachers: differing views of behavior of cleft palate children. Cleft Palate J. 15: 360, 1978.
21. Schneiderman, C.R. and Auer, K.E. The behavior of the child with cleft lip and palate as perceived by parents and teachers. Cleft Palate J. 21: 224, 1984.
22. Chetpakdeechit, W., Hallberg, U., Hagberg, C., et al. Social life aspects of young adults with cleft lip and palate: grounded theory approach. Acta Odontol Scand. 67: 122, 2009.
23. Richman, C.L., Clark, M.L. and Brown, K.P. General and specific self-esteem in late adolescent students: race x gender x SES effects. Adolescence. 20: 556, 1985.
24. Bernstein, N.R. and Kapp, K. Adolescents with cleft palate: body-image and psychosocial problems. Psychosomatics. 22: 697, 1981.
25. Broder, H. and Strauss, R.P. Self-concept of early primary school age children with visible or invisible defects. Cleft Palate J. 26: 114, 1989.
26. Young, S.E., Purcell, A.A. and Ballard, K.J. Expressive language skills in Chinese Singaporean preschoolers with nonsyndromic cleft lip and/or palate. Int J Pediatr Otorhinolaryngol. 74: 456, 2010.
27. Collett, B.R., Stott-Miller, M., Kapp-Simon, K.A., et al. Reading in children with orofacial clefts versus controls. J Pediatr Psychol. 35: 199, 2010.
28. Peter, J.P., Chinsky, R.R. and Fisher, M.J. Sociological aspects of cleft palate adults: IV. Social integration. Cleft Palate J. 12: 304, 1975.
29. Peter, J.P., Chinsky, R.R. and Fisher, M.J. Sociological aspects of cleft palate adults. III. Vocational and economic aspects. Cleft Palate J. 12: 193, 1975.
30. Peter, J.P. and Chinsky, R.R. Sociological aspects of cleft palate Adults: I. Marriage. Cleft Palate J. 11: 295, 1974.
31. Ramstad, T., Ottem, E. and Shaw, W.C. Psychosocial adjustment in Norwegian adults who had undergone standardised treatment of complete cleft lip and palate. I. Education, employment and marriage. Scand J Plast Reconstr Surg Hand Surg. 29: 251, 1995.
32. Hunt, O., Burden, D., Hepper, P., et al. The psychosocial effects of cleft lip and palate: a systematic review. Eur J Orthod. 27: 274, 2005.

(continued from page 28)

3. J Crossley, G.H., B Jolly, Assessing health professionals. Medical Education, 2002. 36: p. 800-804.
4. McGuire, C., Perspectives in Assessment. In: Gonella JS, Hojat M, Erdmann JB, Veloski JJ, eds. Assessment Measures in Medical School, Residency, and Practice: The Connections. New York, NK: Springer Publishing Company, 1993: p. 3-16.
5. Norcini JJ, H.E., Hawkins RE, Evaluation Challenges in an Era of Outcomes-Based Education. In: Holmboe ES, Hawkins RE, eds. Practical Guide to the Evaluation of Clinical Competence. Philadelphia, PA: Mosby Elsevier, 2008: p. 1-9.
6. van der Vleuten CPM, N.D., How can we test clinical reasoning? The Lancet, 1995. 345: p. 1032-1034.
7. Vleuten, C.v.d., The assessment of professional competence: developments, research and practical implications. Advances in Health Science Education, 1996. 1: p. 41-67.
8. Joint Committee on Standards for Educational and Psychological Testing of the American Education Research Association, t.A.P.A., and the National Council on Measurement in Education., Standards for Educational and Psychological Testing. Second Ed. 1999. American Educational Research Association. Washington, DC. , 1999.
9. Schwartz RW, D.M., Young B, Nash PP, Witte FM, Griffen WO., Undergraduate Surgical Education for the Twenty-first Century. Annals of Surgery, 1992. 216(6): p. 639 - 647.
10. Program Director Guide to the Common Program Requirements [cited 2013 May 1]; Available from: http://www.acgme.org/acgmeweb/tabid/237/GraduateMedicalEducation/InstitutionalReview/ProgramDirectorGuidetotheCommonProgramRequi.aspx.
11. Leach, D., The ACGME Competencies: Substance or Form? Journal of the American College of Surgeons, 2001. 192(3): p. 396 - 398.
12. G, M., The assessment of clinical skills / competence / performance. Academic Medicine, 1990. 65(Supplement): p. S63-S67.
13. SM Downing, T.H., Validity threats: overcoming interference with proposed interpretations of assessment data. Medical Education, 2004. 38: p. 327-333.
14. Murphy KR, D.C., Psychological Testing: Principles and Applications. Fifrth Ed. 2001 Prentice-Hall, Inc. Upper Saddle River, NJ.
15. Newble DI, B.A., Elmslie RG, A comparison of multiple-choice tests and free-response tests in examinations of clinical competence. Medical Education, 1979. 13: p. 263-268.
16. SM, D., Reliability: On the reproducibility of assessment data. Medical Education, 2004. 38: p. 1006-1012.
17. MJ, K., Threats to Score Comparability with Applications to Performance Assessments and Computerized Adaptive Tests. Educational Assessment, 1999. 6(2): p. 73-96.
18. Hawkins RE, S.D., 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: p. 42-59.
19. TM, H., Roles and Importance of Validity Studies in Test Development. In: Downing SM, Haladyna ™, eds. Handbook of Test Development. Mahwah, NJ: Lawrence Erlbaum Associates, 2006: p. 739-753.
20. G, P., Against Multiple Choice Questions. Medical Teacher, 1979. 1(2): p. 84-86.
21. Elstein, A., Beyond Multiple-choice Questions and Essays. The Need for a New Way to Assess Clinical Competence. Academic Medicine, 1993. 68(4): p. 244-249.
22. Schuwirth LW, v.d.V.C., Donkers HH, A closer look at cueing effects in multiple-choice questions. Medical Education, 1996. 30: p. 44-49.
23. EA Farmer, G.P., A practical guide to assessing clinical decision-making skills using the key features approach. Medical Education, 2005. 39: p. 1188-1194.
24. Schuwirth LW, v.d.V.C., Different written assessment methods: what can be said about their strengths and weaknesses? Medical Education, 2004. 38: p. 974-979.
25. Newble DI, B.A., Elmslie RG, Assessing clinical competence at the undergraduate level. Medical Education, 1992. 26: p. 504-511.
26. Swanson DB, N.G., Linn RL, Performance-Based Assessment: Lessons from the Health Professions. Educational Researcher, 1995. 24(5): p. 5-11, 35.