Doi: 10.21608/JHPEI.2024.266664.1019

REVIEW ARTICLE **Open Access**

Standard Setting Methods for the Assessment of Knowledge and Skills in Medical Education

Muhammad Saaiq[1]

[1] **FCPS (Plastic Surgery), FCPS (Surgery), MHPE (Master Health Professions Education), MBBS (Khyber).**
Consultant Plastic and Reconstructive surgeon,
National Institute of Rehabilitation Medicine (NIRM), Islamabad, Pakistan

Abstract
Background

The Problem and Gap: Standard setting for assessments in medical education continues to be a less well-understood area. The new educationists are often faced with dilemmas regarding the best choice of the method for standard setting.

The Hook: Clear understanding of the science that underpins the standard setting methods will guide the educationists and faculty members on how to establish standard setting for a given assessment tool.

Methods

A systematic search was carried out for the relevant publications. The search engines used included the PubMed, Google Scholar and ERIC. Additional manual search was undertaken to avoid missing any relevant articles. The standard setting methods employed for knowledge, skills and performance were critically assessed.

Results

The literature found is presented in the PRISMA flow chart. The resultant synthesized literature provided the theoretical background of the standard setting with practical demonstration of the process of standard setting using the Angoff' method for MCQs and borderline regression method for OSCEs.

Conclusion

There is no universal consensus regarding the best method for standard setting of assessments of knowledge and skills in medical education at present. For written assessments, the preferred standard setting methods are those that are item-focused or test-centered. For skills assessment, the borderline regression method is considered to be the most prudent one.

Keywords
Standard setting; Assessment; Assessment of knowledge; Assessment of skills; Assessment tools; Medical education

Received: 2024-01-30 Accepted: 2024-02-26

Published Online: April 2024

Introduction

Assessment of students' achievement of learning is a crucial component of any medical education curriculum. On one hand, assessment drives learning on part of the students; whereas for the medical education institutions, the assessment is mandatory for various other indications such as issuing a valid degree or license. A standard refers to the cutoff level or the minimum pass level of the student's achievement of learning. It is not an arbitrary measure, but rather a measure based on sound scientific methodology [1-3].

There is no gold standard in the standard stetting techniques; however, any method that we choose should be defensible and based on informed judgments. The method should be feasible with respect to the available staff and resources. It should be acceptable to the examinees, examiners and the involved institutions. It should be easy to explain and implement [4-6].

This review aims to outline a comprehensive overview of the various standard setting methods employed for assessments in medical education. It will enhance the understanding of the methodology involved in the standard setting methods for various assessment tools.

How to cite this article
Saaiq, M. "Standard Setting Methods for the Assessment of Knowledge and Skills in Medical Education." J Health Prof Edu Innov, Vol. 1, no. 2 April 2024, pp 14-20.
Doi 10.21608/JHPEI.2024.266664.1019

Correspondence Address
Dr Muhammad Saaiq,
Consultant Plastic and Surgeon Medical educationist,
National Institute of Rehabilitation Medicine (NIRM), Street No.9, G-8/2, Islamabad,
Email muhammadsaaiq5@gmail.com , +923355411583
(ORCID Id 0000-0003-1714-0491)

Methods

Search strategy

A methodological and systematic search strategy was employed to find answer(s) to the following research question: What are the standard setting methods for the assessment of knowledge and skills in medical education.? The key terms were defined and the databases of PubMed, Google Scholar, ERIC were systematically searched.

Key terms used

The following search string was used for the PubMed literature search: (Standard setting methods*) AND (Assessment tools* OR knowledge* OR skill*) AND (medical education).

Review period: Jan 01, 1950 to Dec 31, 2023.

Inclusion criteria:

All publications relevant to the research question, published between 1950 and 2021 were included.

Exclusion criteria:

The following publications/ literature were excluded: Gray literature; Abstracts and conference proceedings; Non-English language literature.

Studies selected:

Relevant articles were selected through the phases of identification, screening, eligibility determination and final inclusion in the synthesis. (Figure-1)

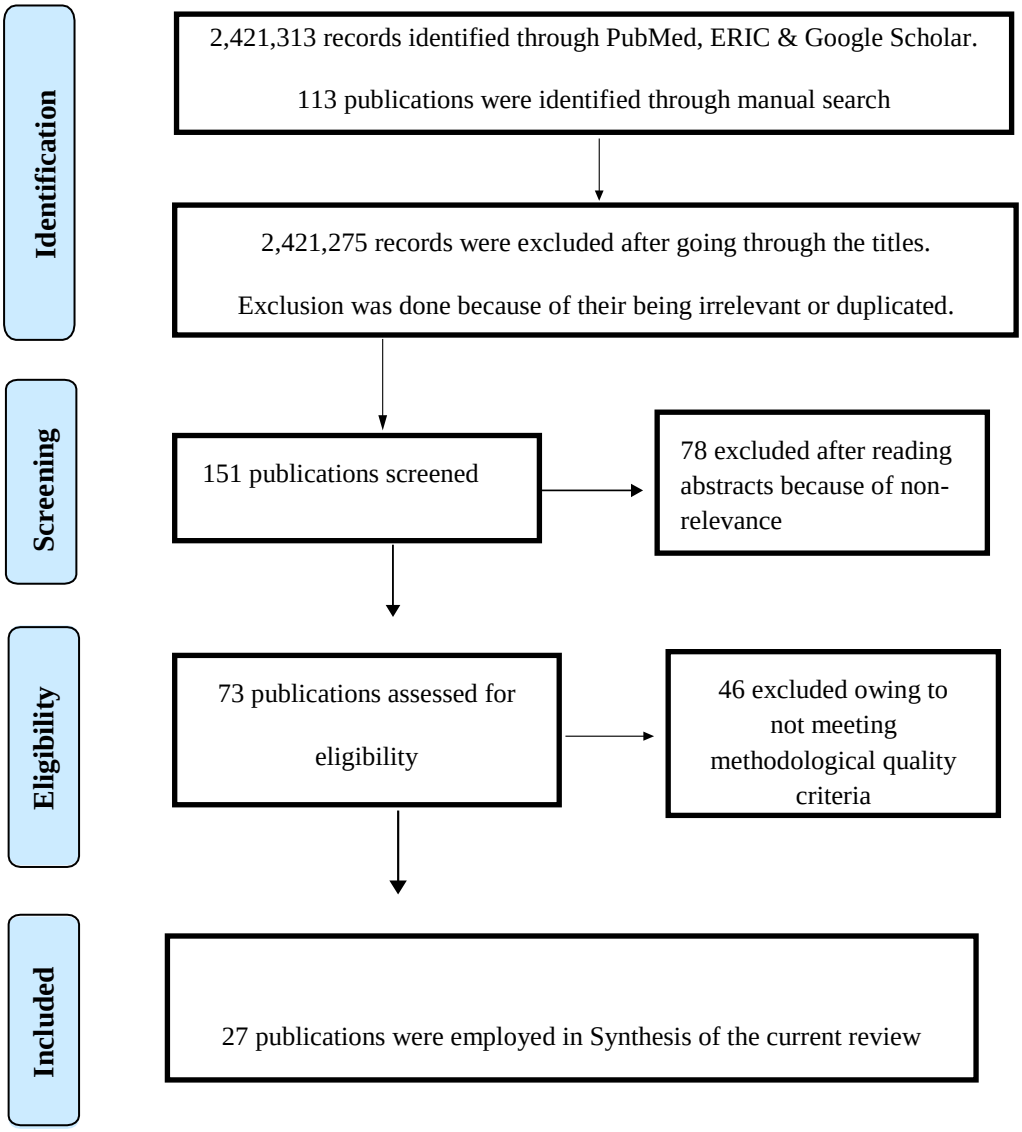

Fig.1: PRISMA Flow Chart: Search results based on PRISMA categorization

Data analysis

The finally included articles were critically analyzed to find answer(s) to the research question.

Results

The literature review identified a heterogenous plethora of standard setting methods for the written assessments. Angoff's and modified Angoff's methods [7-8] were the most commonly employed absolute methods in this regard. These were followed by the Ebel's and modified Ebel's methods [5]. Borderline regression (BLR) method was found to be reported as a reliable and defensible method for the standard setting of skills [9]. A variety of standard setting methods were used for the assessment of Performance. e.g., Anchored rating scales and relative (Norm-based reference) methods [5, 9].

Table (1): Various standard setting methods employed for knowledge and skills performance

Table 1: Various standard setting methods employed for knowledge and skills performance.	
A-**For the Knowledge domain (as is typically assessed in multiple choice questions (MCQs), the following standard setting methods have been employed: [4-7].**	
1	Angoff's and modified Angoff's methods: These methods are the most commonly employed absolute methods. They are reasonably defensible, credible and reliable. Major limitation is the resource intensity involved in the process of standard setting [8,9].
2	Ebel's and modified Ebel's methods: These methods are especially useful when developing and managing MCQs-banks. The modified version helps to ensure that the content is balanced (in terms of essential, important and acceptable) and so is the item construct (in terms of easy, moderate and difficult) [5].
3	Standard error of measurement (SEM) method [2, 7].
4	Wijnen method: The mean minus two times the SEM is taken as the cutoff score in the Wijnen method. It is a relative method. It is easy to use and is affordable. Main limitation is that a fixed percentage of candidates (16%) will fail. Additionally, a large reference group (over 100 candidates) is required [2, 7].
5	Cohen's method: It is a relative method. It employs the top scoring candidates as the reference point. e.g., The 95th percentile score is used, and an arbitrary percentage (say upper 60%) of these determine the passing score. It is easy and affordable [10].
6	Holistic (absolute) method [2, 8, 12].
7	Ad hoc absolute method: (An arbitrary percentage of mastery of domains is defined as a cutoff score, usually by tradition or convention. It is also called the traditional pre-fixed standard setting method [2, 7, 11].
8	Ad hoc compromise method: (any of the relative procedures combined with the (ad hoc) absolute methods [2, 7, 11].
9	Miscellaneous other methods such as the: i. Hofstee method, ii. De Gruijter method, iii. Bookmark method, iv. Beuk method, v. Nedelsky method, and vi. Jaegar method [2,3, 7, 11-13].
B-For Skills (as can be typically assessed in OSCE), the following standard setting methods have been employed: [9, 11,14,19-21].	
1	Borderline regression (BLR) method. It is reliable and defensible.
2	Borderline group method. Entails both scoring of performance as well as a holistic judgment by the experts.
3	Contrasting groups method. Similar to the borderline group method.
4	Miscellaneous other methods: i. Up-and-down method ii. Fixed percentage method iii. Relative Wijnen method iv. Relative 95th percentile method v. Holistic (absolute) method vi. Hofstee method vii. Angoff's method viii. Modified Angoff's method ix. Ebel's method x. Modified Ebel's method xi. Bookmark method

C-For the assessment of Performance (for instance in workplace-based assessments), the following standard setting methods have been employed: [5, 9, 11, 14, 18].	
1	Anchored rating scales: These are similar to Likert scales with descriptors at various points. For instance, they may span from very poor to excellent for performing clinical examination of a standardized patient.
2	Relative (Norm Based Reference) methods. Typically employed in entry tests for admissions.
3	Miscellaneous other methods: i. Borderline regression (BLR) method ii. Borderline group method iii. Contrasting groups method iv. Fixed percentage method v. Hofstee method vi. Modified Angoff's method vii. Angoff's method viii. Ebel's method ix. Modified Ebel's method x. Bookmark method [5, 9, 11, 14, 18].

Discussion

Although there is no consensus about the gold standard, the preferred standard setting methods for written assessments are those which are item-focused (i.e., test-centered). Hence in Table 1, the first five are more frequently employed for standard setting of MCQs. The remaining methods have also been employed in the past for the said purpose by several eminent authors [5,7,13-18].

Although there exists no consensus about the gold standard, there is growing body of evidence in favor of BLR as being the superior standard setting method for skills assessment. The other methods given in the list have been used in the past for standard setting of skills. [9, 11,14,19-21].

The Angoff's method involves a group of subject matter experts (usually 6-8 in number) who make probable estimates about the proportion of borderline examinees who would correctly answer the MCQs. The experts perform this estimation process for each of the included questions. These proportions are then summed for all the expert judges. The median sum of item proportions across judges is taken as the cut-score on the examination [2, 8, 11].

Over the years, several modifications of the original Angoff's method have been introduced. For instance, the experts are allowed to review and refine their judgments after holding a general discussion among themselves regarding the included items. Alternatively, they may be provided with detailed item analysis after the first round of ratings and hence make more informed decisions in their revised judgments [11,17, 22].

The Angoff's method and its modifications have stood the test of time. They continue to be the most favored methods for standard setting in MCQs. They allow the experts to employ their anecdotal experience to judge the items. The modifications allow for more informed judgments.

Following are the inherent limitations of this method: (1) the experts have to define minimal proficiency and (2) consistently estimate the proportions of minimally proficient candidates who would correctly answer each test item [12,18].

The following is a practical case scenario of how to set the standards for single best answer MCQs by employing the Angoff's method:

Table (2): Application of Angoff's method to a five item MCQs test, where eight judges involved in the process of standard setting.

	Standard setters/ Judges							
Questions	1	2	3	4	5	6	7	8
MCQ-1	0.50	0.40	0.20	0.10	0.20	0.30	0.30	0.20
MCQ-2	0.20	0.50	0.05	0.20	0.30	0.40	0.40	0.10
MCQ-3	0.30	0.35	0.15	0.25	0.35	0.10	0.20	0.15
MCQ-4	0.60	0.60	0.30	0.35	0.4	0.50	0.50	0.20
MCQ-5	0.40	0.75	0.60	0.4	0.50	0.60	0.35	0.50
Overall Cut-score	2 (or 2/5)	2.60 (or 3/5)	1.3 (or 1/5)	1.3 (or 1/5)	1.75 (or 2/5)	1.9 (or 2/5)	1.75 (or 2/5)	1.15 (or 1/5)
Overall Cut-score for the five MCQs (all Judges' Mean) = 2+2.60+1.3+1.3+1.75+1.9+1.75+1.15=13.75/8=**1.71/5 or 2/5**								

Table No. 2 shows application of the Angoff's method to a test which comprises of five MCQs. There is a panel of eight experts who are performing the standard setting. For each MCQ, each judge has estimated the proportion of borderline candidates who would probably respond correctly.

The following steps are followed:

Meeting of the assessors' group is convened and orientation is given. The group identifies the features of an imaginary group of borderline candidates who stand 50/50 chance of passing.

The group members assume their roles. One member serves as the convener and documents the points. The other members serve as the judges/examiners.

In the first round, the first item (MCQ) is read. Each judge independently estimates the proportion of the borderline group who are expected to respond correctly. For instance, a given judge may estimate 40% of borderline candidates correctly answering a specific MCQ. The procedure is repeated for all MCQs. All the estimates are recorded.

The first round is followed by an open discussion on the MCQs and the awarded ratings.

The discussion is followed by the second round of final ratings in which the judges make revised and final estimates. The procedure is repeated for all MCQs. All these final estimates are recorded.

The final estimates of judges for each MCQ are collected and averaged to determine the cut off score for that item. The process is repeated for all the items.

The sum of final estimates for all MCQs is calculated and divided by the number of judges. The resultant figure is divided by the total number of items (MCQs), thus constituting the passing score for the entire test.

Following is the case scenario of how is the process and application of borderline regression method (BLR) for OSCEs:

The BLR method has gained popularity for the standard setting in OSCE assessments. This method is less time consuming and is based on actual observation, rather than on a hypothetical borderline candidate's performance. As the name implies, this method uses linear regression modelling to predict the cut off score. The pass score for a given OSCE station is obtained by regressing the candidate scores (e.g., checklist scores) onto the global ratings. (i.e., fail = 1, borderline = 2, pass = 3, good = 4, very good = 5).

In BLR method, the assessors are asked to complete the mark sheet for each candidate, on each OSCE station. The mark sheet has two components, namely the checklist score and the global rating score. The checklist score (usually scored from 1 to 10 marks) may have previously been standard set by employing some other method. The global rating score (usually 1-5) is based on the subjective opinions of the assessors about the individual performance of the candidates on each station. The checklist scores are plotted on Y-axis whereas the global rating scores are plotted on X-axis. The resultant borderline regression graph helps to determine the pass-score for the given OSCE station. The process is carried out for each station. The median score value is chosen as the cut off score on the examination [11, 20, 23-27].

Table (3): Application of the BLR method for OSCEs

Candidates/ Examinees	**Checklist Score** Maximum Marks=10 Cut off Marks=6	**Global rating Score** Maximum Marks=5 Cut off Marks=3
Examinee No. 1	8	5
Examinee No. 2	9	5
Examinee No. 3	7	4
Examinee No. 4	6	3
Examinee No. 5	8	4
Examinee No. 6	6	2
Examinee No. 7	8	4
Examinee No. 8	4	1
Examinee No. 9	5	2
Examinee No. 10	6	2

Following is the application example of borderline regression method:

Table no. 3 shows application of the BLR method for OSCEs. The mark sheet for one OSCE station is given where ten examinees have been assessed. The following steps are followed:

a) Selection of judges who are experts in the content area.

b) Orientation of judges with respect to the content and their tasks.

c) Defining the borderline examinees.

d) Making judgments of the performance of examinees on each OSCE station:

- ❖ The checklist score is marked.
- ❖ The global rating score is given.
- ❖ The global scores are statistically regressed against the checklist scores by plotting the global rating on X-axis and the checklist score on Y-axis (Figure 2)

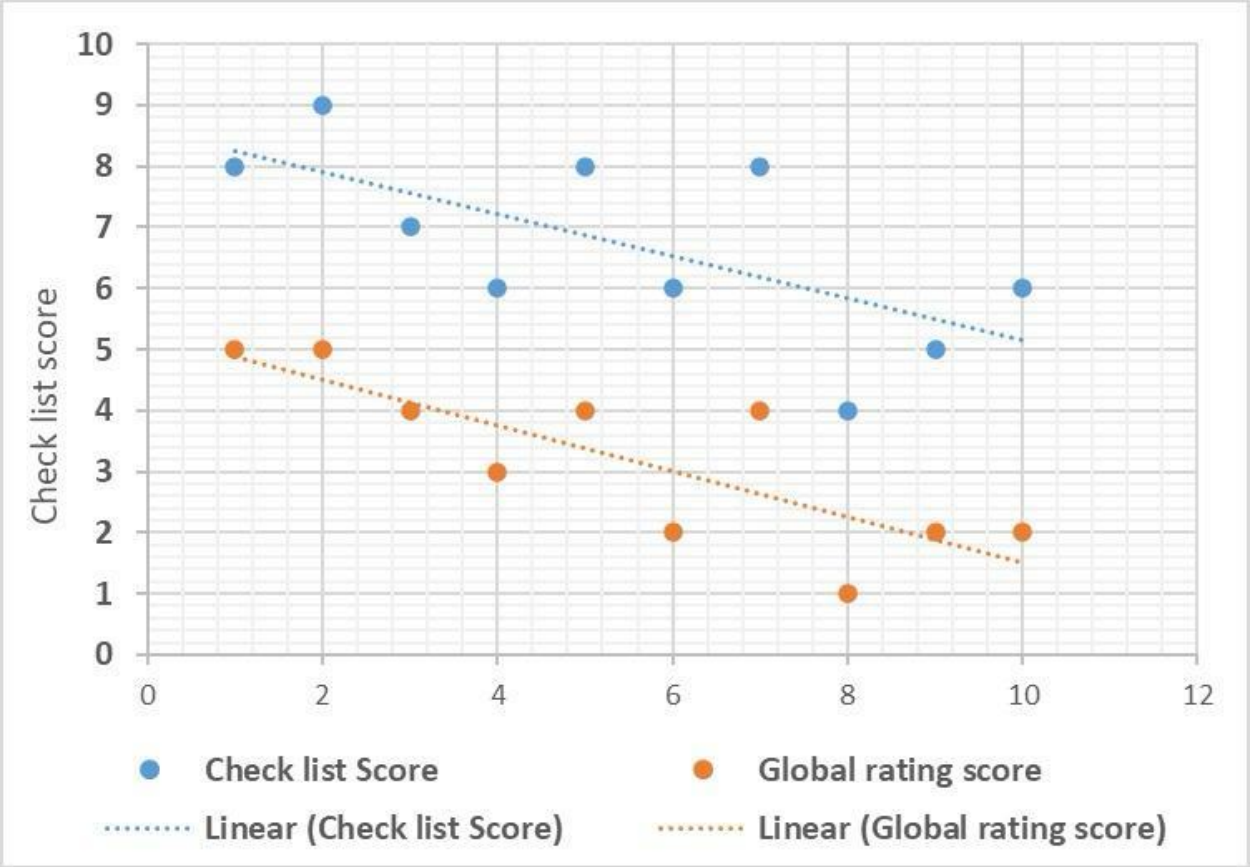

Fig. 2: Scatter diagram for the data displayed in the table 2

e) Setting the pass score: The cut off score is measured using a linear equation by employing the midpoint of the global rating scale against the borderline group(s) scores [11, 20, 23-27].

Conclusion

Currently, there is no universal consensus regarding the best method for standard setting of assessments of knowledge and skills in medical education. For written assessments, the preferred standard setting methods are those that are item-focused or test-centered. For skills assessment, the borderline regression method is considered to be the most prudent one.

Availability of data and material

Not applicable.

Conflict of interest statement

None declared. The author has no financial and personal relationships with any organization that could create a conflict of interest with any material presented in the manuscript.

Financial support/ funding

None declared. There has been no funding or financial support involved in this study.

Author's contributions

MS designed and wrote the manuscript. He performed critical analysis and approved the manuscript.

References

1. Norcini J, McKinley DW. Concepts in assessment including standard setting. In: Dent JA, Harden RM, Hunt D. (Eds.). A practical guide for medical teachers. 5th ed: 2017;252-259. Edinburgh: Elsevier.
2. Bandaranayake RC. Setting and maintaining standards in multiple choice examinations: AMEE Guide No. 37. Med Teach. 2008;30(9-10):836-45.
DOI: 10.1080/01421590802402247.
3. Ben-David MF. AMEE Guide No. 18: Standard setting in student assessment, Medical Teacher 2000, 22:2, 120-130, DOI: 10.1080/01421590078526
4. Hays R, Gupta TS, Veitch J. The practical value of the standard error of measurement in borderline pass/fail decisions. Med Educ. 2008;42(8):810-5. DOI: 10.1111/j.1365-2923.2008.03103.x.
5. A PMETB guide to good practice. Transparent standard setting in professional assessments. In Developing and maintaining an assessment system. London: Postgraduate Medical Education & Training Board. 2007: 13-20. London: Postgraduate Medical Education Training Board (PMETB).
6. vander Vleuten CPM, Cohen-Schotanus J. Standard setting. In: Patel N, Chan LK. (Eds). Assessment in medical and health sciences education. 2009: 63-71. Hong Kong: Institute of medical and health sciences education, The University of Hong Kong.

7. McKinley DW, Norcini JJ. How to set standards on performance-based examinations: AMEE Guide No. 85. Med Teach. 2014;36(2):97-110. DOI: 10.3109/0142159X.2013.853119.
8. Angoff WH. Scales, norms, and equivalent scores. In R.L., Thorndike (Ed.). Educational measurement. 2nd ed, 1971: 508–600. Washington, DC: American Council on Education.
9. Hambleton RK, Plake BS. Using an extended Angoff procedure to set standards on complex performance assessments. Applied Measurement in Education 1995, 8: 41–55. DOI: 10.1207/s15324818ame0801_4
10. Cohen-Schotanus J, van der Vleuten CP. A standard setting method with the best performing students as point of reference: practical and affordable. Med Teach. 2010;32(2):154-60. DOI: 10.3109/01421590903196979.
11. De Champlain AF. Standard setting methods in medical education. In: Swanwick T. (Ed.). Understanding medical education: Evidence, theory and practice. 2nd ed. 2014: 305-316. London, England: Association for the Study of Medical Education (ASME). Wiley-Blackwell Publishing Ltd.
12. Jaeger RM. Certification of student competence. In: Linn RL. (Ed.). Educational measurement. 3rd ed, 1989: 485–514. New York, NY: American Council on Education and Macmillan.
13. De Champlain AF. Setting and maintaining standards in multiple-choice examinations: guide supplement 37.2 - viewpoint. Med Teach. 2010;32(5):436-7. DOI: 10.3109/01421591003677939.
14. Holsgrove G, Kausar SA. Quality assurance, standard setting and item banking in professional examinations. Karachi: College of Physicians and Surgeon Pakistan. 2004.
15. Hussein A, Abdelkhalek N, Hamdy H. Setting and maintaining standards in multiple choice examinations: Guide supplement 37.3--practical application. Med Teach. 2010;32(7):610-2. DOI: 10.3109/01421591003730977.
16. Puryer J, O'Sullivan D. An introduction to standard setting methods in dentistry. Br Dent J. 2015;219(7):355-8. DOI: 10.1038/sj.bdj.2015.755.
17. Ricker KL. Setting cut-scores: a critical review of the Angoff and modified Angoff methods. The Alberta Journal of Educational Research 2006, 52, 53–64.
18. Zieky MJ. So much has changed: How the setting of cut-scores has evolved since the 1980s. In: Cizek GJ. (Ed.). Setting performance standards. 2001: 19–51. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
19. Kaufman DM, Mann KV, Muijtjens AM, van der Vleuten CP. A comparison of standard-setting procedures for an OSCE in undergraduate medical education. Acad Med. 2000;75(3):267-71. DOI: 10.1097/00001888-200003000-00018.
20. Livingstone SA, Zieky MJ. Passing scores: A manual for setting standards of performance on educational and occupational tests. 1982. Princeton, NJ: Educational Testing Service.
21. Van Nijlen D, Janssen R. Modeling judgments in the Angoff and Contrasting-groups method of standard setting. Journal of Educational Measurement 2008, 45(1): 45–63.
22. Jalili M, Hejri SM, Norcini JJ. Comparison of two methods of standard setting: the performance of the three-level Angoff method. Med Educ. 2011 ;45(12):1199-208. DOI: 10.1111/j.1365-2923.2011.04073.x.
23. Cizek GJ, Bunch MB. (2007). Standard Setting: A guide to establishing and evaluating performance standards on test. 2007. Thousand Oaks, CA: Sage Publications.
24. Dauphinee WD, Blackmore DE, Smee S, Rothman AI, Reznick R. Using the Judgments of Physician Examiners in Setting the Standards for a National Multi-center High Stakes OSCE. Adv Health Sci Educ Theory Pract. 1997;2(3):201-211. DOI: 10.1023/A:1009768127620.
25. Kramer A, Muijtjens A, Jansen K, Dusman H, Tan L, van der Vleuten C. Comparison of a rational and an empirical standard setting procedure for an OSCE. Objective structured clinical examinations. Med Educ. 2003;37(2):132-9. DOI: 10.1046/j.1365-2923.2003.01429.x. Erratum in: Med Educ. 2003;37(6):574. PMID: 12558884.
26. Liu M, Liu KM. Setting pass scores for clinical skills assessment. Kaohsiung J Med Sci. 2008;24(12):656-63. DOI: 10.1016/S1607-551X(09)70032-4.
27. Wood TJ, Humphrey-Murto SM, Norman GR. Standard setting in a small scale OSCE: a comparison of the Modified Borderline-Group Method and the Borderline Regression Method. Adv Health Sci Educ Theory Pract. 2006;11(2):115-22. DOI: 10.1007/s10459-005-7853-1.