The Correlation of Acceptability Index (AI) between Medical teacher and Borderline group of Fourth-year Medical Student at Radiology Department

Kanyarat Katanyoo M.D.[1*]
Atchima Cholpaisal M.D.[1]
Phensri Sirikunakorn M.D. [1]
Chiroj Soorapanth M.D.[2]

[1] Department of Radiology, Faculty of Medicine Vajira Hospital, Navamindradhiraj University, Bangkok, Thailand
[2] Department of Orthopedics, Faculty of Medicine Vajira Hospital, Navamindradhiraj University, Bangkok, Thailand
* Corresponding author, e-mail address: kankatanyoo@edu.vajira.ac.th
Vajira Med J 2015; 59(4): 1-8
http://dx.doi.org/10.14456/vmj.2015.23

Abstract

Objectives: To assess the correlation of "acceptability index" (AI) for Nedelsky method from medical teachers with AI from actual borderline medical students and difficulty index (DI).

Method: A prospective study was conducted in the academic year 2012. Eighty fourth-year medical students from 3 groups of rotations in radiology were invited to define AI by Nedelsky method for 100 multiple-choice questions (MCQs). AI of borderline medical students in 3 groups was assessed their correlation with AI from medical teachers. Additionally, DI from item analysis was also considered.

Results: Nineteen borderline medical students were identified. The correlation coefficients (r) of AI from medical teachers and borderline medical students were 0.23 ($p<0.001$), 0.01 ($p=0.93$) and 0.12 ($p=0.10$), and correspond to the correlation values of AI from medical teachers and DI of 0.24 ($p=0.001$), 0.06 ($p=0.455$) and 0.23 ($p=0.003$) in groups 1, 2 and 3, respectively. While the correlation coefficient (r) between AI from borderline medical students in groups 1, 2 and 3 and DI omit: in each their group (unnecessary) were 0.45 ($p<0.001$), 0.61 ($p<0.001$) and 0.53 ($p<0.001$), respectively.

Conclusions: The correlation of AI from medical teachers with borderline medical students and DI was rather poor whereas AI from borderline medical students showed fairly good associations with DI.

Keywords: Acceptability index, Nedelsky method, borderline students

ความสัมพันธ์ของค่า Acceptability index ระหว่างอาจารย์แพทย์และนักศึกษาแพทย์ชั้นปี 4 กลุ่มคาบเส้นที่ภาควิชารังสีวิทยา

กันยรัตน์ กตัญญู พ.บ., วว.รังสีรักษา[1*]
เพ็ญศรี ศิริคุณากร พ.บ., วว. รังสีวิทยาทั่วไป[1]
อัจจิมา ชลไพศาล พ.บ., วว.รังสีรักษา[1]
จิโรจน์ สูรพันธุ์ พ.บ., วว. ออร์โธปิดิกส์[2]

[1] ภาควิชารังสีวิทยา คณะแพทยศาสตร์วชิรพยาบาล มหาวิทยาลัยนวมินทราธิราช กรุงเทพฯ ประเทศไทย

[2] ภาควิชาออร์โธปิดิกส์ คณะแพทยศาสตร์วชิรพยาบาล มหาวิทยาลัยนวมินทราธิราช กรุงเทพฯ ประเทศไทย

* ผู้ติดต่อ , อีเมล kankatanyoo@edu.vajira.ac.th

Vajira Med J 2015; 59(4): 1-8
http://dx.doi.org/10.14456/vmj.2015.23

บทคัดย่อ

วัตถุประสงค์: เพื่อศึกษาความสัมพันธ์ของค่า acceptability index (AI) โดยวิธีของ Nedelsky ที่ได้จากอาจารย์แพทย์เปรียบเทียบกับนักศึกษาแพทย์กลุ่มคาบเส้น รวมถึง difficulty index (DI)

วิธีดำเนินการวิจัย: เก็บข้อมูลไปข้างหน้าในนักศึกษาแพทย์ชั้นปีที่ 4 ปีการศึกษา 2555 จำนวน 84 คน ซึ่งแบ่งเป็น 3 กลุ่มในระหว่างผ่านการเรียนที่ภาควิชารังสีวิทยา โดยได้ให้นักศึกษาแพทย์ได้ทำค่า AI ข้อสอบปรนัยจำนวน 100 ข้อด้วยวิธีของ Nedelsky จากนั้นได้รวบรวมค่า AI ที่ได้จากนักศึกษาแพทย์เฉพาะกลุ่มคาบเส้นเพื่อวิเคราะห์ความสัมพันธ์เทียบกับ AI ที่ได้จากอาจารย์แพทย์ รวมถึงค่า DI จากการวิเคราะห์ในข้อสอบแต่ละข้อร่วมด้วย

ผลการวิจัย: นักศึกษาแพทย์กลุ่มคาบเส้นมีจำนวนทั้งสิ้น 19 คน ความสัมพันธ์ระหว่างค่า AI ที่ได้จากอาจารย์แพทย์เทียบกับนักศึกษาแพทย์กลุ่มคาบเส้นในกลุ่มที่ 1,2 และ 3 เท่ากับ 0.23 ($p<0.001$), 0.01 ($p=0.93$) และ 0.12 ($p=0.10$) ตามลำดับ ซึ่งสอดคล้องกับความสัมพันธ์ระหว่างค่า AI ที่ได้จากอาจารย์แพทย์เทียบกับค่า DI คือ 0.24 ($p=0.001$), 0.06 ($p=0.455$) และ 0.23 ($p=0.003$) ในขณะที่ความสัมพันธ์ระหว่างค่า AI ที่ได้จากนักศึกษาแพทย์กลุ่มคาบเส้นในกลุ่มที่ 1,2 และ 3 กับค่า DI เท่ากับ 0.45($p<0.001$), 0.61($p<0.001$) และ 0.53 ($p<0.001$) ตามลำดับ

สรุป: ความสัมพันธ์ของค่า AI ระหว่างนักศึกษาแพทย์กลุ่มคาบเส้นและอาจารย์แพทย์ค่อนข้างไม่ดี ในขณะที่ค่า AI ของนักศึกษาแพทย์กลุ่มคาบเส้นกับค่า DI อยู่ในระดับดี

Introduction

One of the assessment tools for radiology course of our undergraduate medical curriculum is multiple choice questions (MCQs). We have used criteria-reference method to determine passing score. As we know, criteria-referenced standards have two models.[1]The first is a test-centered model. This model has several methods including Nedelsky[2], Angoff[3], and Ebel[4]. The similarity of all methods is that panelists assume themselves to be a borderline student. In our department, we used Nedelsky method which panelists judge the distractors which borderline students should know in each item. The probability of passing rate in each item is the reciprocal of remaining options which borderline students are not sure whether it is the correct answer or not. For 5 options, the score will range from 0.2-1. We define this score for each item "acceptability index" (AI). If AI is closed to 1, it means that item is easy for borderline students. The summation of AI for all items in a test is minimal passing level (MPL). The difficulty of using this method is how much every panelist knows characteristics of borderline students in the same way. However, the understanding and judgment from panelists may be dissimilar with real borderline students. For another model, examinee-centered models, for example the Borderline-Group method[1], require the panelists looking for the real borderline students and used their median score as passing score. However, this method can be used in case of assembly scores.[5] Moreover, that passing score is for one test, so this method cannot apply to each item.

After the end of the examination process, the quality of each item is displayed in item analysis. One of the most meaningful data reflecting on real difficulty of each item is difficulty index (DI). This index is the percentage of examinees providing right answers. The value is ranged from 0-1, and has the concept of interpretation resemblance to AI. For the easy item , DI is approximate to 1. Thus in one MCQ, there are AI from panelists before examination and DI from item analysis after examination.

From the problem of panelists to simulate themselves as borderline students, the issue about the reliability of their AI with actual borderline students is questionable. The purpose of this study is to evaluate the correlation of AI from medical teachers with AI from a real borderline group of fourth-year medical students by Nedelsky method. DI in each item after finishing examination is analyzed to find the relationship of this value with AI from both sources as well.

Methods

After an approval from the Ethics Committee for Research involving Human Subjects of the institution, we conducted a prospective descriptive study of 4th-year medical students studying at radiology department in the academic year 2012. In the 4th year, medical students were arranged into three-group rotations. Twenty-seven medical students in each group assessed AI by marking the distractor answers at their own question-sheet during 100-MCQ items by themselves. The criteria for choosing borderline student was lower one-third in each group, but still passed the examination. When the examination process finished, real borderline students were recognized. Then AI values of real borderline students in each three group were collected and assessed their correlation of AI from medical teachers with the same method (the Nedelsky method). Additionally, difficulty index (DI) values from item analysis were also considered. Minimal passing level (MPL) deriving from the summation of AI of borderline medical students, medical teachers and DI values were identified.

Statistical analyses

Data were analyzed using SPSS statistical software version 11.5 (SPSS Inc.,Chicago,IL). The correlations between AI values of real borderline medical students versus medical teachers and also DI in each item were evaluated with Pearson Correlation (r) or Kendall's tau as appropriate. A two-sided p-value < 0.05 was determined as statistical significance.

Results

In the academic year 2012, the total number of 4th-year borderline medical students at the department of radiology was 19 from a total of 80 students. For 19 students, mean age was 22.6± 0.9 years. The number of them in group 1, 2 and 3 was 7, 6 and 6, respectively. The distribution of AI from medical teachers, AI from real borderline medical students and DI from100 MCQs in three groups of examination were different and shown in figures 1, 2 and 3, respectively. For AI from medical teachers, there was a right-skewed distribution in the second and third groups. It implied that the majority of 100 AI values from medical teachers were low values. In contrast to AI values from medical teachers, AI from real borderline medical students and DI had the distribution as left-skewed distribution. Additionally, this distinction was a more obvious notice in the third group.

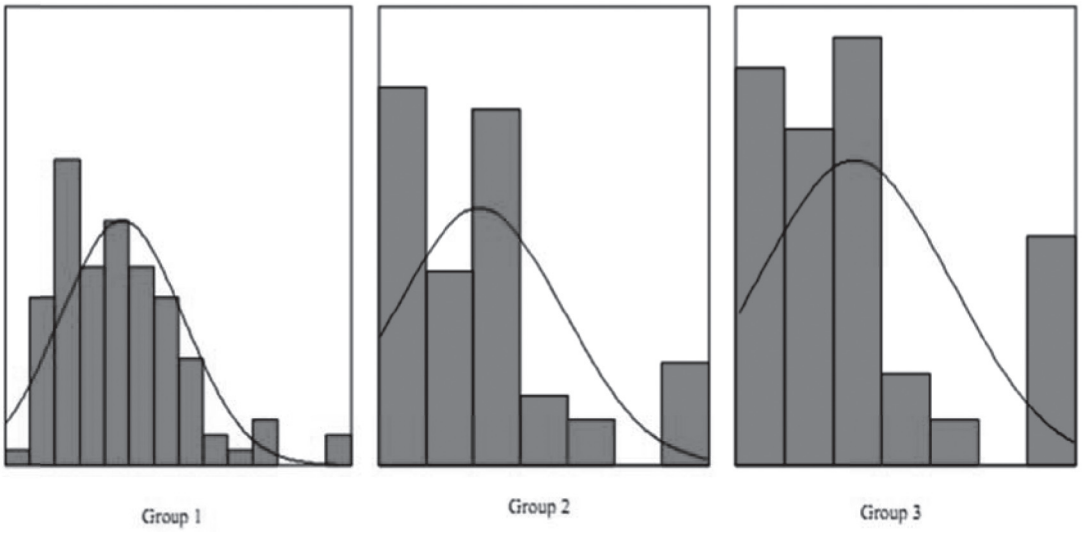

Figure 1: Acceptability index (AI) from medical teachers (x axis = AI value from 0-1, y axis = frequency from 0-100)

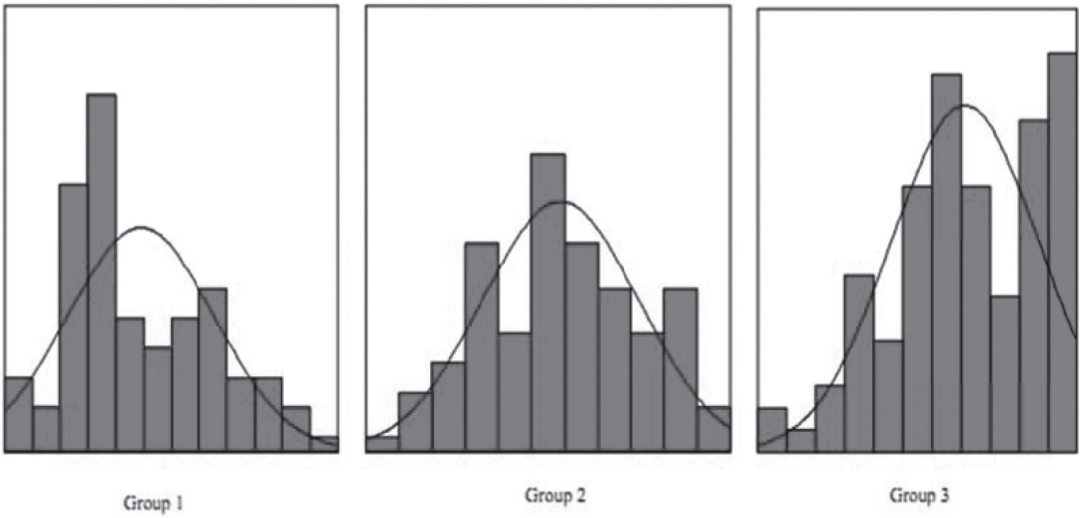

Figure 2: Acceptability index (AI) from real borderline medical students (x axis = AI value from 0-1, y axis = frequency from 0-100)

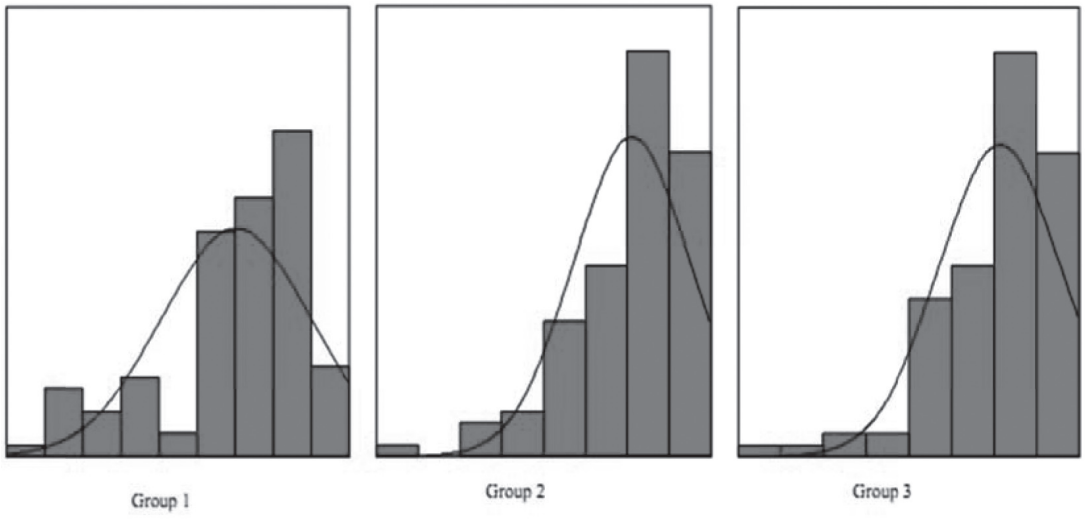

Figure 3: Difficulty index (DI) (x axis = DI value from 0-1, y axis = frequency from 0-100)

According to the non-normal distribution of some AI and DI values, we used Kendall's tau for the statistical test report. The correlations (r) between AI from medical teachers and real borderline medical students were 0.01-0.23. There was only one group, the first group, which had a good correlation between AI from medical teachers and real borderline medical students (r= 0.23, p-value<0.001). The second and third groups had a poor value of this relationship. For the association of AI from medical teachers and DI, the range of r was 0.06-0.24. There were good values in the first and third groups with statistically significant. The last interesting analysis was a correlation between AI from real borderline students and DI. All groups demonstrated the highest correlation values (0.45-0.61) with p-value < 0.001. All results were shown in table 1. When AI of two sources and DI from 100 MCQs were summed to be MPLs, we found that the lowest value of MPL was found in a medical teacher. While MPL values from real borderline students and DI were approximate values. All results were presented in table 2.

Discussion

Standard setting is a necessary process for guaranteeing the competency of examinees which examiners can accept at least. In medical education perspective, this setting is not only to demonstrate passing or failure rate but also to make institution take responsibility for a society which medical graduates provide service. Therefore, this process should be performed deliberately by

Table 1:

Correlation (r) between AI from medical teachers Vs AI from real borderline students Vs DI

Group (n)	AI medical teachers Vs AI borderline students	AI medical teachers Vs DI	AI borderline students Vs DI
1 (7)	0.23 (p<0.001)	0.24 (p=0.001)	0.45 (p<0.001)
2 (6)	0.01 (p=0.93)	0.06 (p=0.455)	0.61 (p<0.001)
3 (6)	0.12 (p=0.10)	0.23 (p=0.003)	0.53 (p<0.001)

Table 2:

MPL from each group

Group (n)	Medical teachers	Borderline students	DI
1 (7)	45.1	58.7	69.2
2 (6)	45.4	71.2	82.8
3 (6)	49.2	83.1	79.0

all stakeholders. The criteria-referenced standards have been recommended to be more appropriate method than norm-referenced standards for the aforementioned purpose.[6] The Nedelsky method that is originated about 60 years ago is one of standard setting by criteria-referenced.[2]

In this study, the graph of AI or MPL in each item from medical teachers demonstrated right-skewed distribution. This was referred that medical teachers tend to underestimate the borderline students. As a result of their estimation, MPL for a test for all groups was lower than 50 out of 100. In 1976, Andrew[7] studied two different standard settings of criteria-referenced between the Nedelsky and the Ebel method. They found the considerable difference in the passing score from two methods. The lower score, around 20%, was obtained with the Nedelsky method. Harasym compared the passing rate of second-year medical student between using the Nedelsky and Angoff method.[8] Almost all the examinees (99%) passed the test when the Nedelsky method was used with more passing rate than the Angoff method about 10%. Moreover, Chang reviewed 10 studies with 40 comparisons between those two methods. They found that the Angoff method produced higher MPL than the Nedelsky method 80% of these comparisons.[9] The reason of low cut-off score when the Nedelsky method was applied might be from the ordinal scale of this method. For 5 options on one item, AI from one judge has been fixed just 5 values including 0.2, 0.25, 0.33, 0.5 and 1. Four values of these are equal or lower than 0.5 whereas only one value is more than 0.5. Therefore, more probability of panelists would describe AI for borderline student yields the values of 0.2-0.5. While AI of the Angoff or Ebel method is not discrete figure as the Nedelsky method. It is a real continuous scale. Both methods may provide the higher score than the Nedelsky method.

According to low AI values from medical teachers, the correlation of them was poor when compared with AI from real borderline medical students. This result finally reflected MPL of the test. Real borderline students were given the difference of MPL from their teachers with higher than 10 points. These differences increased in the second and third groups. The factor which should be taken into consideration, besides the validity of conceptualization of borderline students, was the variation of knowledge of them in the different time. Due to the educational process of block system in our institution, 4thyear medical students who were passed at radiology rotations before other rotations would have a chance to have knowledge less than other groups later. Generally, the knowledge about radiology is integrated with almost other departments. For example, chest radiography in pulmonary tuberculosis was taught in medicine rotation or plain abdominal radiography was taught in surgery rotation. Therefore the last group of students might be more able to comprehend course description of radiology. For estimation of AI value in different time of examination, we should concern about the different level of knowledge from their previous learning as well. The same item should not have the same value of AI if it is used in different time.

Determination of qualified panelists is both important and difficult. Bandaranayake mentioned that process of choosing panelists can make the explicit process of standard setting more.[6] The

significant attribute of panelists is being specialists in each field of examination. Additionally, these specialists should have an intimate relationship with borderline students because they have to simulate themselves as a real borderline student to assess AI value. However, this simulation is the most difficult process in standard setting. It needs enough time for training and has the agreement of concept concerning borderline students between specialists. This point is the limitation for our panelists. Although they are both specialists and medical teachers, that cannot guarantee the validity and knowledge estimation of actual borderline students.[10] Moreover, Kane suggested that the specialists should come from different contexts including medical teachers, physicians and young doctors for the inclusive viewpoint.[11] This idea sounds good but rather impracticable in a real situation. At the present time there is no assessment tool to measure the quality of each specialist or medical teachers to be panelists. Thus the internal validity of all methods of the test-centered model is still being a problem. In this study, we analyzed DI value in each item and compared with AI value from both medical teachers and borderline students. We found good correlations between DI and AI from borderline students in all groups. It is questionable whether we can use DI for adjustment AI from medical teachers by Nedelsky method before judging the examination. It may be one type of modified Nedelsky method and make a more validity for this method. Furthermore, this concept is also used in one of modified Angoff method.[6] However, the most difficult issue which we should concern is external validity. Whether their radiology score at 4th year is correlated with the ability of interpretation or using radiology for solving problems of their future patients after graduation is questionable. Since there are several contaminated factors which can effect on these results.

In the aspect of reliability, we should realize that there are both intra-rater and inter-rater reliability. For intra-rater reliability, the Nedelsky method had more reliability than the Angoff method.[9] While the Angoff method showed much more consistency than the Nedelsky method regarding inter-rater reliability.[9,12] The proper number of panelist has been concerned because it is affected on inter-rater reliability. In regard to the Angoff method, several studies demonstrated the variation of the proper number of panelists. The optimal number of panelists was ranged from 5 in the study of Livingston and Zieky[1], 12-19.5 in the study of Maurer et al.[13] and 10-15 in the study of Hurtz and Hertz.[14] However, the proper number of the panelist for the Nedelsky method still remains doubtful. In our study, we used average AI values which were determined by 2-3 medical teachers in each item. However, we did not study both intra-rater and inter-rater reliability and this is one of our limitations.

Another limitation of this study is reusing some items, around 50-60% per group, which had the history of good performance of DI and discrimination (r). We found that DI values from reused-questions were increased in some items when compared with the first time of using. All these items were not adjusted AI values from their original determination. Due to our curriculum was arranged into group rotations, the examinations took place more than one time in one academic year. Therefore, it is inconvenient to create all new or unexposed questions for every group. As a consequence of these revealing questions, panelists should be critically considered their previous both AI and DI to form the new AI appropriately in every examination. Moreover, one of the important components in item analysis and need to be concerned is distractor analysis. We can know the performance of each option, especially unchosen options. If examiner did not adjust or change these options, panelists have to keep this data in mind when AI is reconsidered. Otherwise passing score in that examination will be underestimated the difficulty. This factor may be a reason why there are great differences increasingly from group 1 to group 3 between AI from medical teachers and actual borderline medical students in the present study. This issue was quoted by "maintenance of examination standards" from Bandaranayake.[6]That author emphasized the sustaining of standard

setting for all examination in one academic year thoroughly. Kane mentioned that standardization of every test is more important than modification the passing score for each test to hold the accurate criteria[11], however, this recommendation was not followed by the medical teacher.[6] This problem was also occurred in our study because of an uncomfortable feeling of them to adjust reusing items for several times in the same year.

Currently, there is no gold standard and most suitable method for setting a point for minimally qualified examinees so long as there are some subjective concerns. Although the idea and process in each method to find out cut-off point is different, the same thing for all methods is to need adequate time for faculty training. The further question is how long enough training can make sure both validity and reliability of criteria. Performance data of item analysis, especially DI, is questionable to use for judgment. As the statement of Bandaranayake, even though there are still some problems in test-centered methods, no method is worse than that.[6]

Conclusions

The correlations of AI from estimation by medical teachers with borderline examinees and DI values were poor. On the other hand, AI values from real borderline students were fairly good associations with DI values. In addition, those were relevant with MPL which was the lowest value from medical teachers and nearly the same value between real borderline students and DI.

Acknowledgements:

We wish to thank all 4th-year medical students in the academic year 2012 and medical teachers at radiology department for their cooperation. This work is granted by medical research fund of Faculty of Medicine Vajira Hospital, Navamindradhiraj University.

References

1. Livingston SA, Zieky MJ. Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service, 1982.
2. Nedelsky L. Absolute grading standards for objective tests. Educ Psychol Meas 1954; 14: 3–19.
3. Angoff WH. Scales, norms and equivalent scores. In: Thorndike RL, editor. Educational Measurement, 2nd ed. Washington DC: American Council on Education 1971; 508–600.
4. Ebel RL. Essentials of Educational Measurement. Englewood Cliffs, NJ: Prentice-Hall, 1972.
5. Ben-David MF. Standard setting in student assessment: AMEE Guide No. 18 Med Teach 2000; 22: 120–30.
6. Bandaranayake RC. Setting and maintaining standards in multiple choice examinations: AMEE Guide No. 37 Med Teach 2008; 30: 836–45.
7. Andrew BJ, Hecht JT. A preliminary investigation of two procedures for setting examination standards. Educ Psychol Measur 1976; 36: 45-50.
8. Harasym PH. A comparison of the Nedelsky and modified Angoff standard-setting procedure on evaluation outcome. Educ Psychol Meas 1981; 36: 45–50.
9. Chang L. Judgmental item analysis of the Nedelsky and Angoff standard-setting methods. Appl Meas Educ 1999; 12: 151-65.
10. Impara JC, Plake BS. Teachers' ability to estimate item difficulty: A test of assumptions in the Angoff standard setting method. J Educ Meas 1998; 35: 69–81.
11. Kane M. Validating the performance standards associated with passing scores. Rev Educ Res 1994; 64: 425–61.
12. Colton DA, Hecht JT. A preliminary report on a study of three techniques for setting minimum passing scores. Presented at the Annual Meeting of the National Council on Measurement in Education. New York, 1981.
13. Maurer TJ, Alexander RA, Callahan CM, Bailey JJ, Dambrot FH. Methodological and psychometric issues in setting cutoff scores using the Angoff method. Pers Psychol 1991; 44: 235-62.
14. Hurtz GM, Hertz NR. How many raters should be used for establishing cutoff scores with the Angoff method? A generalizability theory study. Educ Psychol Meas 1999; 59: 885-97.