ORIGINAL RESEARCH

Comparison of results between modified-Angoff and bookmark methods for estimating cut score of the Korean medical licensing examination

Mikyoung Yim

Research and Development Division, Korea Health Personnel Licensing Examination Institute, Seoul, Korea

Purpose: The purpose of this study was to apply alternative standard setting methods for the Korean Medical Licensing Examination (KMLE), a criterion-referenced written examination, and to compare them to the conventional cut score used on the KMLE.
Methods: The process and results of criterion-referenced standard settings (i.e., the modified-Angoff and bookmark methods) were evaluated. The ratio of passing and failing examinees determined using these alternative standard setting methods was compared to the results of the conventional criteria. Additionally, the external, internal and procedural evaluation of these methods were reviewed.
Results: The modified-Angoff method yielded the highest cut score, followed sequentially by the conventional method and the bookmark method. The classification agreement between the modified-Angoff and bookmark methods was 0.720 measured by Cohen's κ coefficient. The intra-panelist classification consistency of modified-Angoff method was higher than bookmark method. However, the inter-panelist classification consistency was vice versa. The standard setting panelists' survey results showed that the procedures of both methods were satisfactory, but panelists had more confidence in the results of the modified-Angoff method.
Conclusion: The modified-Angoff method showed results that were more similar to those of the conventional method. Both new methods showed very high concordance with the conventional method, as well as with each other. The modified-Angoff method was considered feasible for adoption on the KMLE. The standard setting panelists responded positively to the modified-Angoff method in terms of its practical applicability, despite certain advantages of the bookmark method.

Key Words: Standard setting, Modified-Angoff, Bookmark method, Medical licensing examination

Introduction

Criterion-referenced assessment evaluates an individual's achievement compared to the goals and criteria presented in the curriculum. Unlike norm-referenced evaluation, a qualification or licensing test should consider specific criteria to be the most important assessment parameters [1]. However, several national licensing or qualifying exams in Korea use a predetermined cut score, according to which successful candidates must answer 60% of total questions correctly. This cut score is used for licensing examinations in the health professions, including medicine, dentistry, and nursing even though it is arbitrary and has no theoretical or empirical basis. A reasonable approach must be

Received: January 18, 2018 • Revised: April 20, 2018 • Accepted: September 11, 2018
Corresponding Author: Mikyoung Yim (https://orcid.org/0000-0001-5048-0477)
Research and Development Division, Korea Health Personnel Licensing Examination Institute, Jayangro 45, Gwangjin-gu, Seoul 05103, Korea
Tel: +82.2.2087.8818 Fax: +82.2087.8885 email: mkyim@kuksiwon.or.kr

Korean J Med Educ 2018 Dec; 30(4): 347-357.
https://doi.org/10.3946/kjme.2018.110
eISSN: 2005-7288

identified for determining the criteria used to identify passing candidates based on an assessment of the necessary competence for professional practice and the minimum ability level for which licensure is appropriate. Since 2010, a panel-based standard setting method has been implemented for the clinical performance test that is part of the Korean Medical Licensing Examination (KMLE) [2]. Subsequently, standard setting has become an essential part of performance or skill tests in other professions such as dentistry and nursing [3,4]. However, empirical studies of these issues have rarely been carried out for written exams although theoretical suggestions have been made [5,6].

Criterion-referenced and test-centered standard setting methods are suitable for written tests [7]. Angoff method is predetermined criterion referenced method and test centered method. Modified-Angoff method allows panelists given information such as test result and other panelists' rating result to discuss the cut score. So Modified-Angoff method is the most common method used for licensure and certification in the professions achievement tests [8]. In the Angoff method, the panelists examine each test item and estimates the probability that a minimally competent person will correctly answer the item on the test. Panelists should be able to define and operationalize the concepts of 'minimum ability' and 'borderline group' according to the purposes of the examination [9]. The Angoff method is easy to understand and easy to apply. However, assuming the minimum competency is the cognitive burden of the standard setting panel.

The bookmark method is a standard setting method that was proposed by Mitzel et al. [10] in 1996, in which the cut score is calculated by sorting the items in order of difficulty. In order to use the bookmark method, it is necessary to create an ordered item booklet (OIB) that is arranged in order of item difficulty. Using the item response theory (IRT), the bookmark method has the accuracy of measurement. It relieves the cognitive burden of standard setting panel. In particular, the Angoff rating increases the burden of the standard setting panel and to impair the accuracy of the measurement in tests with a large number of items [11]. Therefore, bookmark method is often used as a way to overcome the disadvantages of the Angoff method. Indeed, Angoff based method were replaced by bookmark method to establish cut score on the National Assessment of Educational Progress since 2005 [12]. That study's findings suggest that bookmark method have comparable reliability, resulting cut scores, and panelists' evaluations to Angoff, even have an advantage of shorter in duration and less costly [12]. The national academic achievement evaluation in Korea also use the bookmark method to establish the standards. Therefore, many studies still compare the two methods and explore the applicability [13-16].

Since the standard setting is a decision making process, the validity of the criterion setting is evaluated by the consistency of the rating and by how well the process is performed in accordance with the principle. Kane [17] suggested validity evaluation of standard setting based on external, internal and procedural criteria. It is important to check that the stated principles were obeyed regarding the standard setting procedures and that the activities were conducted consistently.

Therefore, this study applied these most common used two standard setting methods, the modified-Angoff and bookmark methods, to the written test of the KMLE and evaluated their validity. The research questions are as follows. (1) Is there a difference in the cut score and percentage of students categorized by the cut score calculated by the each standard setting method? (2) Is there any difference in the evaluation results according to external criteria for each standard setting method? (3)

Is there any difference in the evaluation results according to internal criteria for each standard setting method? (4) Is there any difference in the results of procedure-based evaluation for each standard setting method?

Methods

Study was followed four steps such as (1) study object setting, (2) discussion of performance level, (3) implementation of standard setting method, and (4) evaluation.

1. Study objects

1) Examination

The research subject was the KMLE in 2014. The KMLE consists of a total of 400 multiple-choice questions. The predetermined conventional cut score is 240, which is 60% of total score. The total number of examinees was 3,287, and the pass rate was 96.7%.

In bookmark method, difficulty and discrimination were calculated by two-parameter IRT in order to construct the OIB. Data analysis was performed using the Bilog-MG3.0 program (Scientific Software International Inc., Skokie, USA) [18].

2) Standard setting panel

The standard setting panel consisted of 14 medical professors, who were content specialists in medicine. Panelists were required to have experience as national test item development over the past 3 years and at least 3 years of experience in medical education. The panelists were selected in order to balance specializations, to promote understanding of the questions and to be familiar with the academic level of the candidates. Classified by specialty, professors were specialized in internal medicine 28.6%, family medicine 21.4%, pediatrics 28.6%, and preventive medicine 14.3%, and plastic surgery 7.1%. The minor specialty's panelists had extensive experience in medical education and national exams and a knowledge of educational evaluation. All panelists have student education experience in medical college, and 57.1% of panelists have participated in item development of this exam.

In terms of ethical considerations regarding the research participants, they are informed the purpose of the research and voluntarily signed a written informed consent form to participate in the research and all data is anonymous. The standard setting process was carried out as follows.

2. Discussion of performance level

The assumptions of various panelists regarding primary care physicians' competence levels were equalized. Based on Dreyfus's model, consensus was made through discussions on primary care physicians' abilities. Dreyfus proposed a five-level competency model that distinguishes the stages of competence from the novice to expert level (i.e., novice, advanced beginner, competent, proficient, and expert) and Carraccio et al. [19] extended Dreyfus's five steps into six steps and described the specific features of each step. Advanced beginner is able to sort through rules and information to decide what is relevant on the basis of past experience. He or she uses both analytic reasoning and pattern recognition to solve problems and he or she is able to abstract from concrete and specific information to more general aspects of a problem. Competent has more expansive experience tips the balance in clinical reasoning from methodical and analytic to more readily identifiable pattern recognition of common clinical problem presentations. He or she sees the big picture; however, complex or uncommon problems still require reliance on analytic reasoning. Ten Cate et al. [20] expressed the steps proposed by Dreyfus in terms of the medical curriculum. According to Dreyfus's competency curve with medical education,

from novice to competent steps are belong in training course in medical education [20]. The purpose of this test is to distinguish the proficiency of the primary care practitioner. Each stage of competence has several characteristics, which change as the level increases. Panelists discussed the personality of each step and concluded that the minimally competent candidate who would pass a medical licensing examination was not completely competent, but between the advanced beginner and competent stages. Through discussion, the minimum competent person who will pass the medical licensing exam is assumed to be a person who has medical knowledge and is capable of applying to basic and common medical treatment.

3. Implementation of standard setting method

First, the modified-Angoff method was performed. Panelists reviewed each item individually and rated it individually. After individual decision making, panelists modified their decisions in multiple rounds before calculating its final rating. It is desirable to produce a cut score via several iterations [21]. The rating was carried out in three rounds. At each rating stage, panelists recorded the rating results on the provided rating forms, which were collected and reported. Impact information was provided after each rating session, and panelists discussed it. The results of the previous rating session, the resulting pass rate, and the actual item information were given as feedback prior to the next rating session and used as the basis for the discussion.

Secondly, bookmark method was applied. Panelists indicated a point where minimally competent person would pass and fail the prepared OIB. After confirming individual results, the panelists adjusted their scores through group discussions. After confirming the adjusted score and confirming the passing rate, the round ended with the opinion that the score will not be modified anymore through the whole discussion

4. External, Internal and procedural evaluation

External evaluation is conducted by comparing classification results between methods. The classification of successful versus successful examinees by each method was also evaluated using Cohen's κ coefficient of consistency. Internal evaluation investigates methodological consistency, intra-participant consistency, and inter-participant consistency.

Intra-panelist consistency referred to the correspondence of passing and failing rates according to the standards set by a panelist in the first, second, and final rounds of a method. The inter-panelist degree of agreement was defined as the agreement in the passing and failing rates between each pair of the 14 panelists at each rating stage of a method. Classification agreement was calculated using the kappa coefficient, and separate analyses were conducted for the modified-Angoff and the bookmark methods.

The validity of standard setting procedures should be evaluated in accordance with the clarity of understanding, the practicality of the method, the reliability of results, and the procedures. Panelists were surveyed regarding these issues before and after setting the standards. The response scale used a 5-point Likert scale with responses of 'strongly disagree,' 'disagree,' 'neutral,' 'agree,' and 'strongly agree.' This responses were scored from 1 to 5 point in order. The same understanding of the minimum performance ability was used for both methods, and separate evaluations were conducted for the clarity of understanding of each method, the practicality of the method, panelists' confidence in the scores, and the implementation of the procedure. In addition, the two methods were directly compared in terms of their usability and the appropriateness of applying them to the KMLE.

Results

1. Cut score and classification ratio

According to the final decisions, the bookmark method led to a cut score of 230. The conventional cut score was 240, and the modified-Angoff method led to a cut score of 245. The passing rate was 97.6% when the bookmark method was applied, 96.7% when the conventional method was applied, and 95.8% when the modified-Angoff method was applied. These results are shown in Table 1.

2. External evaluation

The κ coefficient of the classification of successful applicants according to the final ratings of the modified-Angoff method and the bookmark method, was 0.720. The consistency of classification between the first rating of the two methods was low 0.499, but it increased to 0.748 for the second round of rating. The coefficient of consistency between the modified-Angoff method and the conventional method was 0.873, and the coefficient of consistency between the bookmark method and the conventional method was 0.840. The modified-Angoff method was similar to the conventional method than the bookmark method (Table 2).

Table 1. Cut Score and Pass Rate of the Modified-Angoff and Bookmark Methods

Method	Round	Cut score	Pass rate (%)
Modified-Angoff	1st	261	91.50
	2nd	246	95.40
	Final	245	95.80
Bookmark	1st	282	78.20
	2nd	234	97.20
	Final	230	97.60

Table 2. Cohen's κ Coefficient Values for the Standard-Setting Methods

Specification		Modified-Angoff		
		1st	2nd	Final
Bookmark	1st	0.499	0.294	0.272
	2nd	0.474	0.748	0.793
	Final	0.420	0.677	0.720

Table 3. Intra-Panelist κ Coefficient Values for Each Method

Panelist	Round	Modified-Angoff		Bookmark	
		1st	2nd	1st	2nd
p1	2nd	0.934		0.127	
	Final	0.934	1.000	0.194	0.772
p2	2nd	0.800		1.000	
	Final	0.800	1.000	0.232	0.232
p3	2nd	0.838		0.020	
	Final	0.859	0.979	0.020	1.000
p4	2nd	0.750		1.000	
	Final	0.750	1.000	0.507	0.507

(Continued to the next page)

Table 3. (Continued)

Panelist	Round	Modified-Angoff		Bookmark	
		1st	2nd	1st	2nd
p5	2nd	0.915		0.840	
	Final	0.982	0.932	0.943	0.896
p6	2nd	0.007		0.776	
	Final	0.007	1.000	0.776	1.000
p7	2nd	0.829		1.000	
	Final	0.829	1.000	1.000	1.000
p8	2nd	0.245		0.130	
	Final	0.245	1.000	0.089	0.802
p9	2nd	0.635		0.858	
	Final	0.635	1.000	1.000	0.858
p10	2nd	0.870		0.557	
	Final	0.870	1.000	0.445	0.852
p11	2nd	0.955		0.230	
	Final	0.955	1.000	0.173	0.840
p12	2nd	0.947		0.718	
	Final	0.947	1.000	0.002	0.004
p13	2nd	0.128		0.386	
	Final	0.128	1.000	0.059	0.227
p14	2nd	0.799		0.386	
	Final	0.799	1.000	0.071	0.266

Table 4. Intra-Panelist κ Coefficient Values for Each Method

	P	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11	p12	p13
Method 1:	p2	0.60												
modified-Angoff	p3	0.47	0.83											
	p4	0.86	0.49	0.38										
	p5	0.29	0.57	0.72	0.23									
	p6	0.68	0.37	0.27	0.82	0.16								
	p7	0.74	0.85	0.69	0.61	0.45	0.46							
	p8	0.93	0.55	0.42	0.92	0.26	0.74	0.68						
	p9	0.47	0.83	1.00	0.38	0.72	0.27	0.69	0.42					
	p10	0.18	0.37	0.48	0.14	0.73	0.10	0.29	0.16	0.48				
	p11	0.86	0.73	0.58	0.72	0.37	0.56	0.87	0.80	0.58	0.23			
	p12	0.47	0.83	1.00	0.38	0.72	0.27	0.69	0.42	1.00	0.48	0.58		
	p13	0.79	0.80	0.64	0.65	0.42	0.50	0.95	0.72	0.64	0.26	0.92	0.64	
	p14	0.89	0.70	0.56	0.75	0.35	0.59	0.84	0.82	0.56	0.22	0.97	0.56	0.90
Method 2:	p2	0.85												
bookmark	p3	0.64	0.78											
	p4	0.86	0.71	0.52										
	p5	0.62	0.76	0.98	0.51									
	p6	0.60	0.74	0.95	0.49	0.97								
	p7	0.54	0.67	0.88	0.43	0.90	0.92							
	p8	0.58	0.72	0.93	0.47	0.95	0.98	0.95						
	p9	0.54	0.67	0.88	0.43	0.90	0.92	1.00	0.95					
	p10	0.64	0.78	1.00	0.52	0.98	0.95	0.88	0.93	0.88				
	p11	0.64	0.78	1.00	0.52	0.98	0.95	0.88	0.93	0.88	1.00			
	p12	0.11	0.15	0.22	0.08	0.23	0.24	0.27	0.25	0.27	0.22	0.22		
	p13	0.56	0.69	0.91	0.46	0.93	0.96	0.97	0.98	0.97	0.91	0.91	0.26	
	p14	0.64	0.78	1.00	0.52	0.98	0.95	0.88	0.93	0.88	1.00	1.00	0.22	0.91

3. Internal evaluation

The intra-panelist degree of agreement was higher for the modified-Angoff method than for the bookmark method. In particular, 12 out of 14 panelists showed an agreement of 1.0 between the second and final ratings. However, for the bookmark method, the score range of one panelist was very large, meaning that the degree of agreement between the scores was low. In contrast, the agreement among the panelists was higher for the bookmark method than for the modified-Angoff method. The κ coefficient over 0.60 means that the degree of agreement is high. High agreement among panelists was frequently observed for the bookmark method (Tables 3, 4).

4. Procedural evaluation

In order to evaluate the standard setting procedure, the following points were assessed: (1) the clarity of understanding of the definition of minimum competence, (2) the clarity of implementation of the standard setting method, (3) the practicality of the standard setting method, (4) confidence in the scores, (5) the implementation of the procedure, and (6) the ease of rating (Table 5).

Understanding the standard setting method not only refers to the degree to which a panelist is familiar with the procedure, but also the panelist's understanding of the underlying principles assumed by the standard setting method, such as the identification of the minimum performance level. Even if the procedure for applying the criteria is simple and clear, the basic assumptions underlying the method could be unclear. In the modified-Angoff method, the understanding of the clarity of the method was low, but the modified-Angoff method had a rather high possibility of using a standard setting method in which ratings are assigned according to guidelines, with a high ease of rating. This does not mean that the standard setting procedure is unclear, but that it is challenging to understand its basic assumptions. The possibility of rating according to guidelines was scored as 3.79 for the modified-Angoff method and 3.64 for the bookmark method.

The modified-Angoff method showed a positive evaluation of the appropriateness of the cut score with mean score of 3.79, which was confirmed by the degree

Table 5. Survey Questions for Each Method

Content	Question	Modified-Angoff	Bookmark	Same
Clarity of understanding of the minimum ability level definition	Clarity of the definition of minimum competence	2.71±0.726	2.71±0.726	
	Ease of assumption of the probability of the minimum competent person responding in a certain way	2.57±0.756	2.57±0.756	
	Usefulness of the achievement level description	2.50±0.855	2.50±0.855	
Clarity of implementation	Pre-education understanding	3.31±0.751	3.79±0.802	
	Pre-education clarity of the task	3.23±0.725	3.64±0.745	
Practicality of standard-setting method	Rating according to guidelines	3.79±0.579	3.64±0.745	
	Appropriateness of application on the Korean Medical Licensing Examination	3.79±0.699	2.21±1.122	
Confidence in scores	Confidence in my cut score	3.79±0.802	3.54±0.776	
	Confidence in the final cut score of the panel	4.00±0.555	3.46±0.660	
Implementation of the procedure	Usefulness of the discussion after round 1	4.29±0.469	4.00±0.877	
	Usefulness of the discussion after round 2	3.62±0.870	3.79±0.975	
Ease of rating	Which standard-setting method was easier to rate?	9 (64.3)	4 (28.6)	1 (7.1)

Data are presented as mean±standard deviation or number (%).

of certainty of each of the panelists and the degree of their assurance in their own cut scores. The confidence in the final score was higher than that of the panelists' own scores, with mean score of 4.00. In contrast, the scores obtained using the bookmark method showed the opposite tendency. The appropriateness of the panelists' own cut score was scored as 3.54, but the appropriateness of the final panel score was 3.46.

The practicality of implementing the procedure was evaluated based on the panelists' recognition of the usefulness of the discussions between each rating session, the usefulness of the feedback information, and whether an appropriate amount of time was spent. For the modified-Angoff method, since the first round of rating was conducted without the actual item difficulty, the discussion after the first round of rating was considered to be highly useful. The panelists showed a generally positive response to the usefulness of the discussion, the usefulness of the information provided in the discussion, the appropriate of time assignment, and the opportunity to provide comments during the discussion. For both methods, the whole-panel discussions were considered to be more useful than the small-group discussions. The usefulness of the information provided in the discussion was rated as 3.71 for the modified-Angoff method and 3.86 for the bookmark method.

The bookmark method' s mean score was 3.79 for the clarity of pre-comprehension when modified-Angoff's score was 3.31. About clarity of the task, modified-Angoff's score was 3.23, when bookmark's score was 3.64. However the feasibility of the modified-Angoff method's score was rated higher than bookmark's.

In a direct comparison of which method was considered to be easier for rating, only one of the 14 panelists perceived the two methods as being the same, while 64.3% selected the modified-Angoff method and 28.6% selected the bookmark method. If a standard setting method is actually applied to the KMLE, the modified-Angoff method would be preferred over the bookmark method.

Discussion

Based on the above results, the following considerations were discussed. First, the cut score of the modified-Angoff was higher than bookmark's cut score. The cut score of the modified-Angoff method was the highest, while the conventional 60% cut score was higher than that of bookmark method. This is similar to the finding of study of Karantonis and Sireci [22] that the cut score obtained using the bookmark method may be lower than the cut score resulting from other standard setting methods, and agrees with the findings of other studies that the cut score obtained using the bookmark method was lower than that obtained using the modified-Angoff method [13-16].

The passing rate differed by approximately 1% between each method. Thus, neither method is much different from the existing method, so they can feasibly be used as alternatives for determining the cut score. We do not expect that the implementation of a new cut score will cause a major problem in comparison to the existing passing rate if a standard setting method is eventually applied. Although the cut score obtained using either of these new methods would not externally differ from the existing cut score to a notable extent, the intrinsic meaning of the score would be valuable in that it would be derived through a standard setting process conducted by professionals, instead of being an arbitrary score.

Through three rounds, it is observed that the classification agreement of both methods improves. The first round match of each method was only 0.499, but the final result increased to 0.720. Both methods showed

high classification agreement even when the existing method was applied. The final classification agreement between the modified-Angoff method and the existing criteria was the highest, 0.873, while the agreement between the bookmark method and the existing method was 0.840. Therefore, the modified-Angoff method led to results that were more similar to those obtained using the conventional method than those obtained using the bookmark method. Nonetheless, the agreement coefficients confirmed that the two new methods showed very high agreement with the existing method, and the agreement between the new methods was also high.

Second, the intra-panelist classification consistency for the modified-Angoff method was higher than for the bookmark method. However, the inter-panelist classification consistency for the bookmark method was higher than for the modified-Angoff method. For the modified-Angoff method, the intra-panelist agreement was high, but the inter-panelist agreement was low.

In the modified-Angoff method, each panelist did not significantly change his/her values, but the results of the rounds as a whole showed gradual changes due to convergence towards the mean. In contrast, the ratings obtained using the bookmark method changed greatly due to changes in the bookmark position in the order of items. When the bookmark positions of the panelists are adjusted, the inter-panelist agreement is improved. As a result, the effect of the convergence of the evaluation results of the individual panelists into one value was larger for the bookmark method.

Third, the procedural assessment validated the process of applying new methods, with several implications. The huge knowledge of medicine makes it difficult to identify the concepts of minimum competence as a physician and minimum competency holder and to derive the corresponding level of achievement. Standard setting panelists should have more discussion, education, and experience. In both methods, the implementation of the procedure was evaluated as good. The modified-Angoff method received higher ratings than the bookmark method for the appropriateness of the cut score. The discussion and coordination process was very important to the panelists. Therefore, at least three rounds of implementation are required, and sufficient time allocation is recommended to ensure a thorough discussion and to enable suitable adjustments to be made.

Fourth, the panelists felt that the modified-Angoff method was more applicable, and they were more confident in that method. Multiple difficulties in applying the bookmark method were reported. Panelists experienced cognitive confusion about arranging the items in order of difficulty. Because of the difficulty of making a single break point in a test in which various subjects were combined, the applicability of the bookmark method was evaluated negatively. However, the bookmark method, which does not require all items to be examined, requires less time for implementation, evaluation, and discussion, making it efficient and convergent.

Although the convergence of the bookmark method's panelists was better than modified-Angoff method's, the panel consisting of medical specialists recognized that the modified-Angoff method was easier, more appropriate, and more convincing. Although methods based on IRT have many advantages in measurement compared with methods based on classical test theory, they are rarely applied because the public understanding of the concepts underlying IRT is poor. In particular, the modified-Angoff method is preferred in general education fields where the standard setting process is carried out by subject content experts, whereas the bookmark method is not widely used [23]. However, when a test contains many items that must be reviewed and has enough examinee for item analysis, the bookmark method may be feasible because it is efficient, less

time-consuming, and converges well, as shown in the present study. The bookmark method is more appropriate when the content of the test satisfies the unidimensionality assumption of IRT. Therefore, the appropriate method may vary depending on the characteristics of the test. The use of a sufficiently empirical method is a prerequisite for successful application, as this process involves considerable cost and time. Kane [17] stated that one can be assured that a rational set of criteria is used when the criteria are applied by a non-prejudiced panel that understands the purpose of the set of criteria. Regardless, the choice and training of the panelists are important. If conformity is satisfied for basic criteria, such as the dimensionality of the test, the type of the items, and the appropriate reliability, the aspects of implementation should be considered, such as the ease of forming the panel, the ease of panel education, and panel members' prior experience. The standard setting method should be applied considering various factors such as the characteristics of the panelists, quantity and quality of items, time, cost, the impacts of the results, and the stakeholders.

Finally, the limitations and suggestions for future research studies are recommended. The IRT is based on the assumption of unidimensionality. However, KMLE written test is not organized by subject knowledge such as internal medicine, surgery, and pediatrics, but is integrated into medicine. In this multi-dimensional knowledge mixed test, constructing OIB analyzed by IRT and finding one cut off point are both a difficulty of applying bookmark method and also limitation of this study.

The number of panelists and their representativeness limit the generalization of the results. Fourteen panelists may be adequate for the recommended number of people on the panel, but the number may not be sufficient because the examination consists of multiple major areas that are not a single subject. It is necessary to compare the results of more panelists or the results of other panelists. As the content of the exam changes, it is also necessary to compare tests conducted in other years.

ORCID:
Mikyoung Yim: https://orcid.org/0000-0001-5048-0477

Acknowledgements: None.

Funding: This research was supported by the Korea Health Personnel Licensing Examination Institute.

Conflicts of interest: No potential conflict of interest relevant to this article was reported.

Author contributions: All work was done by Mikyoung Yim.

References

1. Cizek GJ, Bunch MB. Standard setting: a guide to establishing and evaluating performance standards on tests. Thousand Oaks, USA: Sage Publication Inc.; 2007.
2. Park H. Clinical skills assessment in Korean Medical Licensing Examination. Korean J Med Educ. 2008;20(4): 309-312.
3. Shim JS, Kim YJ, Kim JA, et al. A study of standard setting for Korea dentist clinical skill test: RE02-1411-02. Seoul, Korea: Korea Health Personnel Licensing Examination Institute; 2014.
4. Shin SJ, Kim YK, Suh SR, Jung DY, Kim YJ, Yim MK. A study of adoption of clinical skill test as Korea nurse licensing examination: RE02-1407-05. Seoul, Korea: Korea Health Personnel Licensing Examination institute; 2014.
5. Ahn DS, Im H. Standard setting in student assessment by criterion referenced evaluation. Korean J Med Educ. 2001;13(1):41-45.

6. Lee G. A psychometric approach to setting a passing score on Korean National Medical Licensing Examination. J Educ Eval Health Prof. 2004;1(1):5-14.
7. Jager RM. Certification of student competence. In: Linn RL, ed. Educational Measurement. 3rd ed. New York, USA: American Council on Education; 1989:485-514.
8. Clauser BE, Margolis MJ, Case SM. Testing for licensure and certification in the professions. In: Brennan RL, ed. Educational Measurement. 4th ed. Westport, USA: Praeger Publishers; 2006:701-731.
9. Angoff WH. Scales, norms and equivalent scores. In: Thorndike RL, Angoff WH, American Council on Education, eds. Educational Measurement. 2nd ed. Washington, USA: American Council on Education; 1971:508-600.
10. Mitzel HC, Lewis DM, Patz RJ, Green DR. The bookmark procedure: psychological perspectives. In: Cizek GJ, ed. Setting Performance Standards: Concepts, Methods, and Perspectives. Mahwah, USA: Lawrence Erlbaum Associates; 2001:249-281.
11. Buckendahl CW, Ferdous AA, Gerrow J. Recommending cut scores with a subset of items: an empirical illustration. Pract Assess Res Eval. 2010;15(6):1-10.
12. Peterson CH, Schulz EM, Engelhard Jr G. Reliability and validity of bookmark-based methods for standard setting: comparisons to Angoff-based methods in the National Assessment of Educational Progress. Educ Meas Issues Pract. 2011;30(2):3-14.
13. Kim TE. The comparison of applicability: Angoff and bookmark methods in standard setting [dissertation]. Seoul, Korea: Ewha Womans University; 2004.
14. Kim NJ. The standard setting of clinical performance examination (CPX) by modified-Angoff, bookmark, and item-descriptor matching (IDM) Method [dissertation]. Seoul, Korea: Ewha Womans University; 2010.
15. Çetin S, Gelbal S. A comparison of bookmark and Angoff standard setting methods. Educ Sci Theory Pract. 2013;13(4):2169-2175.
16. Yin P, Schulz EM. A comparison of cut scores and cut score variability from Angoff-based and bookmark-based procedures in standard setting. Paper presented at: the annual meeting of the National Council on Measurement in Education; April 12-14, 2005; Montreal, Canada.
17. Kane MT. So much remains the same: conception and status of validation in standard setting. In: Cizek GJ, ed. Setting Performance Standards: Concepts, Methods, and Perspectives. Mahwah, USA: Lawrence Erlbaum Associates; 2001:53-88.
18. Zimowski MF, Muraki E, Mislevy RJ, Bock RD. BILOG-MG3.0. Skokie, USA: Scientific Software International; 2003.
19. Carraccio CL, Benson BJ, Nixon LJ, Derstine PL. From the educational bench to the clinical bedside: translating the Dreyfus developmental model to the learning of clinical skills. Acad Med. 2008;83(8):761-767.
20. Ten Cate O, Snell L, Carraccio C. Medical competence: the interplay between individual ability and the health care environment. Med Teach. 2010;32(8):669-675.
21. Hambleton RK. Setting performance standards on educational assessments and criteria for evaluating the process. In: Cizek GJ, ed. Setting Performance Standards: Concepts, Methods, and Perspectives. Mahwah, USA: Lawrence Erlbaum Associates; 2001:89-116.
22. Karantonis A, Sireci SG. The bookmark standard-setting method: a literature review. Educ Meas Issues Pract. 2006;25(1):4-12.
23. Jang YS, Seong TJ. The comparison for IRT based standard setting methods: bookmark, IDM and mapmark. J Educ Eval. 2009;22:659-680.