Comparison of validity of Bookmark and Angoff Standard Setting Methods in Medical performance tests

majid yousefi afrashteh (mjduosefi@gmail.com)

Research article

Keywords: standard setting, Angoff, bookmark

Posted Date: February 25th, 2020

DOI: https://doi.org/10.21203/rs.2.24421/v1

Version of Record: A version of this preprint was published on January 2nd, 2021. See the published version at https://doi.org/10.1186/s12909-020-02436-3.

Abstract

Introduction

One of the main processes in evaluating of the students' performance is standard staging to determine the passage for the test. The purpose of this study was to compare the validity of two methods of Angoff and bookmark in standard setting.

Method

Participants included 190 master's students graduated in laboratory sciences since past year. Designed by a group of experts, a performance test with 32 item was used in this study to assess laboratory skills of graduates of medical laboratory sciences. Moreover, two groups of experts voluntarily participated in this study to set the cut-score. To assess the process validity, a 5-item questionnaire was asked from two groups of penists. To investigate the internal validity, the variance of the cut scores determined by the members of the two panels was compared with the F ratio. External validity was assessed by using four indices of correlation test with criterion score.

Results

Comparison of the two methods of Angoff and bookmarking showed that the mean of process validity indices was higher in bookmarking method. In order to assess internal validity, conclusion: Homogeneity of results and co-ordination of judges' scores were considered.

Conclusion

In evaluating of the external validity (concordance of the cut score with the criterion score), All five external validity indices supported the bookmark method.

Background

Student performance assessment is an important part of educational programs. Since learning objectives for university students are set through performance assessment, it is given serious attention by higher education instructors and planners [1, 2]. All educational systems seek to level up the performance of learners to achieve predefined objectives [3]. Setting the passing/failing grades and/or acceptable performance level (or minimum pass level) is the natural and common outcome of Tests, which is important not only for the learners, but also for higher levels including the school, city, state, and country [4]. Nevertheless, setting cut-off point or passing criterion is less regarded as an assessment pillar [2].

Passing standard, passing criterion, or minimum pass level are hypothetical boundary within score range of a test that distinguishes individuals who have achieved mastery level from those who have not [5, 6, and 7]. Standard setting methods are used to determine the cut-score or minimum pass level [8, 9]. Typically, a fixed score, such as 10 or 12, is considered to be the minimum pass level in educational

examination at university level and employment tests [10]. The use of a fixed pass level for all conditions is not fair, due to the effect of such factors as difference in item difficulty, execution of test, level of subjects, and objective of the test. Therefore, educational justice is obtained when the minimum pass level in each test is set based on the conditions of that test [11].

In general, standard-setting methods are either item-centered or person-centered [12]. In item-centered methods (e.g. Angoff method), test content is reviewed by a group of experts and judges; whereas, in person-centered methods (e.g. boundary groups), decision of judges is based on the actual performance of the subjects [13]. According to literature, Angoff [14] and Bookmark [15] are the most common item-centered methods. Angoff method has been proven by many evidence as the most common and best-known standard-setting method [17, 18, and 19]. In the Angoff method, prior to the conduct of the test, a group of experts and judges are asked to review the content of each test question and then predict the likelihood of correct answer to each item by minimally-qualified candidates. Then, the obtained values are discussed to arrive at consensus by all judges over all questions. Finally, the mean of scores given by the judges to all questions is set as the pass standard and cut-score [19]. Nevertheless, this method is associated with difficulties, such as having a long-term procedure and need for an expert group [17, 20]. In addition, ambiguity in the concept of minimally-qualified student is among other limitations of this method [17, 21]. In an attempt to overcome the shortcomings of the Angoff method, researchers proposed a new method which, in addition to being suitable for both multiple-choice and constructed-response questions, reduces the experts' work load, facilitates their decision-making, combines the experts' decisions with measurement models in determining the cut-score, and considers the test content together with the performance level [22]. This method, named Bookmark, was introduced by Mitzel, Lewis, Patz, and Green [15] and quickly welcomed. In this method, the place of each item in the scale is first determined according to difficulty index extracted by item response theory (IRT) and then the items are placed in separate sheets from the easiest one to the most difficult item. The expert panel is then asked to place its bookmark somewhere between the questions, where they believe the probability of giving correct answer by the minimally-qualified subject is 50% or 67%. After items are determined by the expert panel, their difficulty for all judges is extracted and the mean difficulty score is calculated. The cut-score in this method is determined by converting the mean ability score into raw scores. In the second round, data such as passing and failing rates, based on the obtained cut-score, is given to the experts. Based on the feedback, the experts are able to change their bookmarks between the items. In case of any change, the cut-score is determined again and is given to the panel. This process continues until a general consensus over cut-score is achieved [23, 24, and 25].

Several studies have compared the Angoff and Bookmark approaches. Hsieh (2013) used Angoff and Bookmark approaches to assess language proficiency levels of students [26]. Results of this study showed that the mean scores obtained from 32 experts, using these two methods, were different in determining the three final cut-scores. In this study, the strengths and weaknesses of each method were addressed. Buckendahl, Smith, Impara and Plake [27] compared the Angoff and Bookmark methods in standard-setting for a math test containing 69 items. In this study, a group consisting of 23 experts participated. Both methods produced similar cut-scores. However, standard deviation of experts was

lower in the bookmark method. Reckase compared the modified Angoff method and Bookmark method with simulated data in an ideal condition (without judgment and parameter estimation errors) [28]. This study showed a negative bias in the Bookmark method, meaning that the estimated cut-score was always lower than the cut-score hypothesized by the experts. This study also showed that the modified Angoff method had slight bias or was without skew. Olsen and Smith (2008) compared the modified Angoff and Bookmark methods for a home inspection certification, and found that the results were fairly similar [29]. Results of the two methods were also similar in terms of the standard error of judges and initial cut-score. These results were obtained in Schultz(2006)'s study, too [30].

Since the implementation of all methods for standard setting and selecting the best cut-score is not practically possible, adoption of an appropriate method is an important part of test construction. Hambleton and Pitoniak reviewed different comparative studies but did not found any effective and generalizable result [31]. They highlighted the need for further comparative studies. Plake emphasized that the majority of conducted studies in this area have supported on a specific method and many factors, such as their validity, need to be examined [32]. Cizek and Bunch mentioned current methodological problems facing standard-setting methods [24]. According to Cizek, investigation and comparison of validity of different methods can result in identification of the best one [33]. kane (1994) emphasized on the collection of evidence for the evaluation of three types of validity (namely process, internal, and external) to measure the validity of standard-setting methods. Process evidence pertains to the accuracy of execution process and trust of passing and performance grades, internal evidence relates to the degree of agreement between judges, and external validity refers to consistency of the obtained cut-score with an external objective criterion [34].

Given the importance of standard-setting methods in setting passing grade, especially in performance tests, the present study intended to compare two common standard-setting methods, Angoff and Bookmark, by comparing their validity indices. Previous comparisons between these two methods and other standard-setting methods have been theoretical or merely based on the comparison between their passing rate and standard error; whereas, this study tried to make comparisons between the process, internal, and external validity of these two methods to reveal the advantages of each method based on validity indices.

Method

Designed by a group of experts, a performance test was used in this study to assess laboratory skills of graduates of medical laboratory sciences. This test included 32 multiple-choice practical items pertaining to skills needed in different laboratory sections. Participants included 190 master’s students graduated in laboratory sciences since past year. The subjects were selected randomly from 30 laboratories across Tehran City. Moreover, two groups of experts voluntarily participated in this study to set the cut-score. Two groups with 12 and 11 individuals cooperated in the application of the Angoff and Bookmark methods, respectively. The Angoff method was initially explained to the first group and ambiguities were clarified by presenting the features of a minimally-qualified postgraduate. In the next session, test items

were given to them to estimate the likelihood of an answer to each item by a minimally-qualified subject. The experts assigned a likelihood score to each item. Then, the assigned likelihood scores were discussed to arrive at final consensus. The mean score given by the judges was set as the cut-score. The second group received training on Bookmark method in the first session. After that, the items were calibrated, based on the item response theory (IRT), and ordered based on the difficulty parameter. These panelists were then asked to place their bookmarks somewhere between the questions, where they believe the likelihood of giving correct answer by a minimally-qualified subject is 50%. After items were determined by the panelists, their difficulty levels for all judges were extracted and the mean was calculated. The cut-score in this method was determined by converting the obtained average of ability scores into raw scores. In the second round, data such as passing and failing rates, based on the obtained cut-score, was given to the experts. Based on the feedback, the experts were able to change the position of their bookmarks between the items. The new cut-point in this stage was determined and given to the panelists. After the confirmation by all panelists, the final cut-score was set.

To investigate the process validity of the two methods, a 5-item questionnaire was administered to measure the satisfaction of execution accuracy for each method. To investigate the internal validity, standard Deviation of the two methods were compared. To measure the external validity, in addition to the subject's score in the test, the judgment of the employer, or authority of the unit and supervisor about the subject's performance was also considered. This step was taken to find out how much the test matched an external criterion in passing or failing the subjects. The employer checklist consisted eight items with two-choice for accepting or rejecting the subject's performance.

Results

Results included descriptive data, process validity, internal validity, and external validity as follows. A summary of descriptive data for both methods is presented in Table 1.

Table 1
descriptive statistic for Angoff and Bookmark methods

methods	min	max	Cut score	SD	Pass n	Pass rate
Angoff	16	22	17.67	1.72	118	55.70
Bookmark	18	21	18.82	0.98	98	46.20

According to this table, the lowest and highest cut-scores in the Angoff and Bookmark methods were 16 and 22, and 18 and 21, respectively. The obtained cut-scores in these methods were 17.67 and 18.82, respectively. Standard deviation of the Bookmark group (0.98) was lower than that of the Angoff group (1.72), indicating a Homogeneity between panelists in the Bookmark method. The passing rates and ratios in the Angoff and Bookmark methods were 118 and 55.70, and 98 and 46.20, respectively. Thus passing criteria was relatively stricter in the Bookmark method.

Evaluation indices are presented in Table 2 to investigate the process validity.

Table 2
Mean and standard deviation for proses validity indicators of two methods

Evaluation Items	Angoff		Bookmark	
	M	SD	M	SD
1-Is the procedure was well understood?	4.66	0.44	4.52	0.89
2-Is the process of determining the cut score in this case was reasonable and appropriate?	4.13	0.59	4.88	0.40
3-Is working in the panel for members was desirable?	4.50	0.65	4.61	0.58
4-Is the overall process of this method was accurate?	3.88	0.71	5.00	0.00
5-After knowing the cut-off point, Is it true From your point of view?	4.04	0.50	4.89	0.33
Total	4.25	1.24	4.79	1.33

For the Evaluation of standard-setting process, the panelists were asked about six indices about the two methods. According to the Table 2, the satisfaction of the Angoff method was greater only for the first index. The most difference was in the accuracy of the methods, that the mean values of the Bookmark and Angoff methods were 5 and 3.88, respectively. In addition, general mean values of the execution process for the Angoff and Bookmark groups were 4.25 and 4.79, respectively. The mean execution values were desirable in both groups.

To investigate the internal validity, variance of the seted cut-scores by the members of the two panels were compared with F-ratio. Since the variance of Angoff and Bookmark groups were 2.96 and 0.96, the F-ratio was obtained as 3.08, which was greater than the 2.96 for the $F_{(11, 10, 0.05)}$. This finding indicates that the dispersion of scores in the Bookmark panel was significantly lower than in the Angoff panel. Therefore, it can be said that the bookmark panelists performed significantly more coordinated and Homogeneous, resulting in higher validity in this method.

To investigate the external validity, five indices (correlation with Criterion score, specificity, sensitivity, and positive and negative predictive values) were used. The Tetrachoric correlation coefficient was used to investigate the relationship of passing/failing grades of the test with those of the criterion score (employer's assessment). Sensitivity is the probability of an assessment instrument in correct diagnosis of an individual who, according to the employer, has the target condition (in here failing). Specificity is the probability of an assessment instrument in correct diagnosis of individuals who do not have the target condition (in here failing). The positive predictive value of a test indicates the probability of a person who really has the condition to have a positive test result (in here failing). The negative predictive value of a test indicates the probability of a person who really does not have the condition to have a negative test result (in here passing) [35].

According to Table 3, correlation between passing/failing, according to the employer, and passing/failing, according to the test score, was 0.69 in the Angoff method and 0.88 in the Bookmark method. In both method, correlation coefficient at the alpha < 0.05 was considered to be significant. Four other indices also showed that the passing/failing grades in the Bookmark method was practically closer to real passing/failing level set by the employer. In other words, the standard set by the Bookmark method was more consistent with actual performance of the subjects.

Table 3
External validity evaluation indicators for Angoff and Bookmark methods

method	External validity evaluation indicators				
	correlation	spe	sensi	Pos.pre	Neg.pre
Angoff	0.69	0.86	0.81	0.85	0.83
Bookmark	0.88	0.92	0.96	0.90	0.98

Discussion

Standard setting is one of the key issues in performance tests. A comprehensive comparison of the two standard setting methods in the laboratory skills test was the main subject of this study.

Comparison of the two methods of Angoff and bookmark was done in terms of internal validity, external validity, process validity and achievement rate. The agreement between results and scores given by judges was considered in evaluation of the internal consistency. The comparison of the Angoff and Bookmark methods showed that the panelists in the Bookmark method acted with greater uniformity and harmony. This result is consistent with the findings of Buckendahl, Smith, Impara, and Plake [27].

The process validity of the two methods was compared with the five indices evaluated by the panelists. it was found that the mean of validity indices was higher in the Bookmark method; whereas, the mean of better understanding of the procedure was higher in the Angoff method. This result supported the better implementation process in the bookmark method.

In order to evaluate the external validity (agreement of the cutoff score set in this test with the external criterion), the assessments made by the employers were used as a basis for setting the cutoff score. All five external validity indices supported the Bookmark method. In fact, setting the pass/fail scores by the Bookmark method was more similar to the judgment of the employers about the respondents. Olson and Smith [29] and Schultz [30] compared the two methods of Bookmark and Angoff and reported a similarity concerning the standard error of the judges and cutoff score. In these two studies and almost all other similar studies, the validity criteria, especially the external validity, were not addressed. Based on the Kane's [34] recommendation on validation methods, the current study investigated three types of

validities, namely process, internal, and external, and allowed for a more comprehensive comparison of these two methods. In general, the results of the present study support the superiority of the bookmark method over the method of Angoff. In addition to the validity indices, the acceptance rates of the two methods were also compared. The lower achievement rate in the Bookmark method implied the greater difficulty in reaching passing grade set by this method. This finding is consistent with Çetin and Gelbal [4]. This result may be useful for different purposes of performing the tests.

There are several reasons for the higher validity of the bookmark method. The Bookmark method, which was developed to complement the Angoff method, combines the expert decisions with advanced measurement models (IRT). As a more important reason, it focuses on the test content, along with the performance level. Also Since medical tests have clear practical content, in that the content and practice are very close, the cutoff point set by the Bookmark method was more precise and consistent. Moreover, the Bookmark method showed higher external validity than the Angoff method.

This study had some limitations. The first limitation of this study was the unequal number of panelists in two ways. The reason for this was the departure of one of the panelists of the method of the Angoff. The second limitation concerns employers' evaluation of students in external validity. This criterion was measured by an 8 questions that may not be accurate. Also, although the two methods of Angoff and bookmark are the most common Standard setting, it may be better to compare them with other standardization methods such as Borderline regression.

Conclusions

It is recommended that the future interested researchers compare the performance of these methods in different tests and with different samples. It is also recommended to planners of medical education and assessment centers prioritize bookmark in standard setting.

Declarations

Ethics approval and consent to participate

All participants gave informed written consent with the right to withdraw at any time. In the first part of the questionnaire, there was a paragraph introducing the study aim and assuring confidentiality of data by anonymous questionnaires. Participants did not experience any harm and they were allowed to stop their participation during the data collection process. The design and methods were approved by the Committee of the Faculty of Psychology of the University of zanjan (iran).

Competing interests

The authors declare that they have no competing interests

Consent for publication

Not applicable

Funding

Not applicable

Authors' contributions

Not applicable

Acknowledgements

Not applicable

Availability of data and materials

The datasets during and/or analyzed during the current study available from the corresponding author on reasonable request.

Abbreviations

IRT: Item response theory

References

1. Zlatkin-Troitschanskaia O, Shavelson RJ, Pant HA. Assessment of Learning Outcomes in Higher Education. Handbook on Measurement, Assessment, and Evaluation Higher Education. 2018:686-98.
2. Hejri SM, Jalili M. Standard setting in medical education: fundamental concepts and emerging challenges. Medical journal of the Islamic Republic of Iran. 2014;28:34.
3. Aviso KB, Lucas RI, Tapia JF, Promentilla MA, Tan RR. Identifying Key Factors to Learning Process Systems Engineering and Process Integration through DEMATEL. Chemical Engineering Transactions. 2018 Aug 1;70:265-70.
4. Çetin S, Gelbal S. A Comparison of Bookmark and Angoff Standard Setting Methods. Educational Sciences: Theory and Practice. 2013;13(4):2169-75.
5. Kane M. Validating the performance standards associated with passing scores. Rev Educ Res. 1994; 64:425–61. doi:10.2307/1170678.
6. Cusimano MD. Standard setting in medical education. Acad Med 1996; 71:112–20.
7. Elfaki OA, Salih KM. Comparison of Two Standard Setting Methods in a Medical Students MCQs Exam in Internal Medicine. American Journal of Medicine and Medical Sciences. 2015;5(4):164-7.
8. Cizek GJ, Bunch MB. Standard setting: A guide to establishing and evaluating performance standards on tests. SAGE Publications Ltd; 2007.

9. Liu M, Liu KM. Setting pass scores for clinical skills assessment. The Kaohsiung journal of medical sciences. 2008 Dec 31;24(12):656-63.
10. Jalili M, Mortazhejri S. Standard Setting for Objective Structured Clinical Exam Using Four Methods: Pre-fixed score, Angoff, Borderline Regression and Cohen's. Strides Dev Med Educ. 2012; 9 (1) :77-84.
11. Mortaz Hejri S, Jalili M, Labaf A. Setting Standard Threshold Scores for an Objective Structured Clinical Examination using Angoff Method and Assessing the Impact of Reality Chacking and Discussion on Actual Scores. Iranian Journal of Medical Education. 2012; 11 (8) :885-894
12. Liu M, Liu KM. Setting pass scores for clinical skills assessment. Kaohsiung J Med Sci. 2008;24:656–3.
13. Cizek GJ, Bunch MB. Standard setting: A guide to establishing and evaluating performance standards on tests. SAGE Publications Ltd; 2007.
14. Angoff W. Scales, norms, and equivalent scores. Educational Measurement: Theories and applications. 1996;2:121.
15. Mitzel HC, Lewis DM, Patz RJ, Green DR. The bookmark procedure: Psychological perspectives. Setting performance standards: Concepts, methods, and perspectives. 2001:249-81.
16. Smith RW, Davis-Becker SL, O'Leary LS. Combining the best of Two Standard Setting Methods: the Ordered Item Booklet Angoff. Journal of Applied Testing Technology. 2014;15(1):18-26.
17. Boursicot K. Setting Standards in a Professional Higher Education Course: Defining the Concept of the Minimally Competent Student in Performance-Based Assessment at the Level of Graduation from Medical School. Higher Education Quarterly. 2006 Jan 1;60(1):74-90.
18. Talente G, Haist SA, Wilson JF. A model for setting performance standards for standardized patient examinations. Evaluation & the health professions. 2003 Dec 1;26(4):427-46.
19. Norcini JJ. Setting standards on educational tests. Medical education. 2003 May 1;37(5):464-9.
20. Kilminster S, Roberts T. Standard setting for OSCEs: trial of borderline approach. Advances in Health Sciences Education. 2004 Sep 1;9(3):201-9.
21. Elfaki OA, Salih KM. Comparison of Two Standard Setting Methods in a Medical Students MCQs Exam in Internal Medicine. American Journal of Medicine and Medical Sciences. 2015;5(4):164-7.
22. Lin J. The bookmark procedure for setting cut-scores and finalizing performance standards: Strengths and weaknesses. Alberta journal of educational research. 2006 Apr 1;52(1):36.
23. Cizek GJ, Bunch MB, Koons H. Setting performance standards: Contemporary methods. Educational measurement: issues and practice. 2004 Dec 1;23(4):31-.
24. Cizek GJ, Bunch MB. Standard setting: A guide to establishing and evaluating performance standards on tests. SAGE Publications Ltd; 2007.
25. Lewis DM, Mitzel HC, Mercado RL, Schulz EM. The bookmark standard setting procedure. Setting performance standards: Foundations, methods, and innovations. 2012 Mar 22:225-54.
26. Hsieh M. Comparing yes/no Angoff and Bookmark standard setting methods in the context of English assessment. Language Assessment Quarterly. 2013 Jul 1;10(3):331-50.

27. Buckendahl CW, Smith RW, Impara JC, Plake BS. A comparison of Angoff and Bookmark standard setting methods. Journal of Educational Measurement. 2002 Sep 1;39(3):253-63.
28. Reckase MD. A conceptual framework for a psychometric theory for standard setting with examples of its use for evaluating the functioning of two standard setting methods. Educational Measurement: Issues and Practice. 2006 Jun 1;25(2):4-18.
29. Olsen JB, Smith R. Cross validating modified Angoff and Bookmark standard setting for a home inspection certification. Inannual meeting of the National Council on Measurement in Education, New York 2008.
30. Schulz EM. Commentary: A response to Reckase's conceptual framework and examples for evaluating standard setting methods. Educational Measurement: Issues and Practice. 2006 Sep 1;25(3):4-13.
31. Hambleton RK, Pitoniak MJ. Setting performance standards. Educational measurement. 2006;4:433-70.
32. Plake BS. Standard setters: Stand up and take a stand!. Educational Measurement: Issues and Practice. 2008 Mar 1;27(1):3-9.
33. Cizek, G. J. (2006). Standard setting. In Steven M. Downing and Thomas M. Haladyna (Eds.) *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers, pp. 225258.
34. Kane M. Validating the performance standards associated with passing scores. Review of Educational Research. 1994 Sep 1;64(3):425-61.