PJERE, December 2021, Vol. 9, No. 2, pp.65-75

A Comparison of Standard Setting Methods for Setting Cut-Scores for Assessments with Constructed Response Questions

Muhammad Naveed Khalid
Resource Person, Allama Iqbal Open University, Islamabad
Farah Shafiq
Assistant Professor, Division of Education, University of Education, Lahore
Shehzad Ahmed
Assistant Professor, Faculty of Education, University of Okara, Okara

KEY WORDS

Standard setting, Angoff, Norm-reference, Criterion reference, Pass score

ABSTRACT

Standard setting provides a way to define minimal competency for various professional assessments. In the literature, a number of methods are proposed but there are implications for examinees because they can produce varied passing scores. Angoff is a widely applied method in context of educational assessments to define the borderline student that required extensive training of judges and skills to conceptualize minimum proficiency. The Cohen has defined an alternative procedure to overcome the limitations of Angoff. Additionally, we explored the relative method by computing average of score distribution as a point below that mean as the passing mark. Objective of the study was to investigate performance of Angoff with other standard setting procedures to inform future standard setting practices. These methods were applied to various exams having small, medium and large number of students. We found Angoff method produced credible and reliable pass scores and close to the relative method but Cohen and Modified Cohen gave divergent results. We recommend studied standard setting procedures explored further with different formats of assessments having varied sample sizes.

Introduction

The Undergraduate programmes in the School of Dentistry at various universities (Bachelor of Dental Surgery, BSc in Hygiene and Therapy and Dental and Diploma of Hygiene) are required to meet General Dental Council (GDC) Standards in order to ensure that qualifications are registerable with the GDC. These standards served as regulatory tool to ascertain that programmes offered at the university are fit for purpose. GDC

(2012) standards broadly covered many aspects and one of them describes about the assessment of students is:

"Assessment must be fair and undertaken against clear criteria. Standard setting must be employed for summative assessments".
According to Kane (1994) standard is a theoretical border that categorized the students who had the required minimum level of competency or from those who do not have. In other words, it provides the description whether performance of a student for a particular purpose is good enough (Kane, 1999). Standard setting is a judgmental process, in which qualified experts determine "How much is enough" through the established set of activities and they determine a numerical score that corresponds to it (Kane, 2013).

Numerous standard setting methods reported in the literature for clinical and written assessments. Kane stated that there is a no definitive answer that which standard setting method is perfect (Kane, 2013). The appropriateness of method depend on the context and aim of the test because each method has advantages and disadvantages. Therefore, the determination of passing score must demonstrate the transparency, reproducibility, credibility and feasibility (Kaufman, et al, 2000; Wass, et al, 2001). Standard setting methods can be classified into three groups namely compromise, absolute (criterion referenced) and relative (norm referenced) methods (Cizek, 2012).

Relative standards are useful to rank the students and based on actual score distribution. A fixed percentage of students has to fail irrespective of the exam difficulty or proficiency of candidates because cut score are decided in advance (Cohen-Schotanus & Van der Vleuten, 2010). However, these methods are quite straightforward and easy to implement but important aspects namely exam difficulty and ability level of candidates were not taken in to account in defining pass scores. It is reported extensively in the literature that these factors affect the cut scores (McKinley & Norcini, 2013). Therefore, experts recommended the absolute method might be used to judge the competencies related to the health profession (Norcini, 2003). Competency level determined by these methods are not influenced by the performance of a particular group of students (Downing & Yudkowsky, 2009). A panel of experts gave their opinion what would be minimum level of competency they are looking for particular assessment. Methods based on absolute criterion broadly categorized into student or exam centred (Livingstone & Zieky, 1982). Primary focus of exam centric methods is the content of assessment and examples of these standards are Jaeger (1982), Angoff (1971), Nedelsky (1954) and Ebel (1972). On contrary performance of examines is a major concern nor content of test to determine pass score in student centric methods. Borderline regression and contrasting groups are well known methods that belong to this category (Wood et al, 2006).

Standards that take into account the features of above both categories are called compromise methods. Compromise methods such as Cohen (2010) and Hofstee (1983) are extensively studied in the literature.

The field of standard method is not without controversy because research has shown that passing score or cut score depends on a particular procedure (Jaeger, 1989; Zieky, 2001). Kane (1994) stated validation of any standard setting method could not be completely tested . Hence, it is more important related evidences support the credibility of the standards. Downing et al (2006) argued that choice of method is an institutional decision and focus should be on how cut score derived. They suggest reliability, validity, replicability and fairness are important characteristics of the process.

Previous comparison of standard setting methods in the medical education have indicated great variability in cut off scores in OSCEs (Kaufman et al, 2000; Humphrey-Murto et al, 2002) and in MCQs (George et al, 2006; Omer et al, 2015). However, to our knowledge, we have not found any comparative study that examined the application various procedures and their effect on determination of pass score in an examination comprised of multiple short answer questions in the dental education. Therefore, it is critical to explore the consistency of established bench marking procedures and to see their application that comprised of multiple short answer questions. This study compared the outcome of absolute method (modified Angoff), compromise method (Hofstee and Cohen) and relative method (Mean, SD) on examinee performance in constructed response questions (multiple short answers) to judge students competencies in the final, middle and early years of dentistry assessments.

In particular, we investigated to find out the questions (1) what is the credibility of cut scores resulting from absolute, relative and compromised standard setting methods? (2) Do the pass rates and cut score depends on the particular method?

Sample

Student scores were collected across written examinations that comprised of multiple shorts answer questions. Summative assessments were comprised of three papers having 15, 10 and 12 compulsory questions that administered to the students who appeared in the Year 1, Year 3 and Year 5. The questions covered wide range of topics prescribed in the curriculum. The number of participants in the Y1 and Y3 were 81 while 75 students appeared in the Year 5.

Methods of Standard Setting

In the standard setting exercise, a panel of judges participated who were faculty members at the school of dentistry. All participants were experienced and familiar with the school curriculum and taught courses. Two days training was given to the participants about the standard setting and through the consensus, the definition of minimally competent/borderline student was developed. Following the independent and individual ratings from each Judge, panel assembled and discussed the ratings against each question. An opportunity was provided to the judges if they want to revise their ratings. We observed high consistency and reasonable variation across judges in their ratings and found no hawks or doves. We used one normative, one absolute and two compromise standard setting procedures for this comparative study. A brief summary of each method is given below.

The Angoff Method

This approach is frequently used to determine the pass score in educational settings and there are numerous modifications to the original Angoff method (Cizek, 2012). First step is define the characteristics of a hypothetical candidate that would have minimum competency level to be considered as pass. Panelists reviewed each test question and told to indicate the score that a hypothetical candidate may attain.

For SBAs and MCQs - What proportion of minimally competent candidates would answer each question correctly?

For MSAs and OSCEs: Estimate how many marks the minimally competent candidate would obtain on each question/station?

The passing score is derived by taking the average of the ratings for each question across the judges.

Relative/Norm Method (Mean – SD)

The computation of this method depends on the distribution of the test score of students. The boundary of minimal qualified score was derived by subtracting the standard deviation (SD) from the mean of the test scores. In the present study, we used mean minus 1 SD to determine the passing mark. The drawbacks of the relative standards are that a fixed percentage of students bound to fail, cut score are not determined in advance and the performance of students can influence the passing score.

Hofstee Method

The Hofstee method take into account the merits of absolute and relative standard setting methods (De Gruijter, 1985). The panellists do not give score against each item instead raters review the whole assessment to judge the difficulty level and give their opinion about what would be the lower and higher pass level and failure rates for the exam. Cut score is derived by taking the average of estimates across panellists and interaction is found on the students score distribution.

Cohen Method

Cohen is a compromised method in which the students who achieved higher score (the 95th percentile) are used as benchmark and cut score is established by taking 60% of the score (Cohen-Schotanus & Van der Vleuten, 2010). The Cohen method also has elements of both relative and absolute standards. Recently, Taylor (2011) explored whether the underlying assumptions of the Cohen method hold or not. They found 60% of score of the 90^{th} percentile produced consistent results for their historical data rather than the 95^{th} percentile when compared to using a fixed pass mark. In our study, we examined which modification of the Cohen method could be implemented in our context in particular when there was a small sample size.

Results

The descriptive statistics were computed to describe the basic quantitative features of the data using SPSS. There was a good range of marks (shown in table 1) which shows that the assessment discriminates well between the stronger and weaker students. The average score was comparable across the Year's specific papers. We can see from the below figures that approximately 95% of the scores fall within Mean ± 1SD across papers. We also examined whether the distribution of score was normal using the Shapiro-Wilk test of Normality and Q-Q plots. We found the scores were reasonably normally distributed for the majority of the assessments.

Table 1
Descriptive Statistics

	Y5P1	Y5P2	Y5P3	Y1P1	Y1P2	Y1P3	Y3P1	Y3P2	Y3P3
N	76	76	76	80	80	79	80	79	80
Mean	89.8	93.1	93.5	62.6	68.2	62.2	62.9	60.5	62.2
Median	91.0	93.5	95.0	62.0	66.0	61.0	63.0	60.0	63.0
Mode	88.0	83.0	95.0	55.0	64.0	75.0	69.0	60.0	64.0
Std. Deviation	10.8	9.7	8.5	16.4	15.8	17.5	8.6	8.8	7.2
Skewness	-0.8	-0.6	-0.6	-0.3	0.2	0.1	-0.6	-0.6	-0.7
Kurtosis	0.1	0.1	0.7	0.9	-0.4	-0.7	0.6	0.1	3.1
Range	47.0	42.0	43.0	93.0	66.0	70.0	44.0	41.0	49.0
Minimum	61.0	67.0	67.0	7.0	35.0	28.0	33.0	35.0	32.0
Maximum	108.0	109.0	110.0	100.0	101.0	98.0	77.0	76.0	81.0

Cronbach's alpha was computed to estimate the reliability of the scores. The coefficient for internal consistency reliability of each paper shown in the table 2. Addition to that we computed standard error of measurement (SEM) that indicates how much error is associated with observed scores. Smaller value of SEM shows scores are assessed with more precision. Figures showed the error of measurement was small that indicates our assessments had higher reliability.

Table 2
Reliability and Standard Error of Measurement

	Y5P1	Y5P2	Y5P3	Y1P1	Y1P2	Y1P3	Y3P1	Y3P2	Y3P3
Alpha	0.72	0.76	0.77	0.79	0.78	0.82	0.77	0.76	0.77
SEM	4.79	3.91	4.87	7.45	7.18	6.77	4.11	4.31	3.43
G	0.88	0.86	0.86	0.95	0.91	0.86	0.83	0.82	0.85
Phi	0.84	0.83	0.83	0.89	0.89	0.83	0.81	0.80	0.82

We also examined the inter rater reliability that shows how much consensus among the judges. Range of this coefficient indicates how much judges vary in their scores and helpful to flag Hawks and Doves. Generalizabiliy theory offers a way of looking at sources of variability within an exam and determining its reliability. We considered two sources of variability: questions and examiners using single facet crossed design. We computed

variance components to calculate a Generalizabiliy coefficent (G) and Dependability coefficent (Phi). G and Phi coefficients were quite high, shown in the table 2, from the norms which are reported in the literature. We observed high consistency and reasonable variation across judges in their ratings and found no outliers.

The comparison of Pass/Fail percentages of studied standard setting methods given in table 3 and table 4 which were calculated based on the test scores. We found wide variation in passing rates among the procedures. Methods were compared based on the students' percentage who passed or failed the respective exams. The Pass/Fail status of the students varied because procedures yielded different cut scores for each assessment.

For early clinical years (Y1), Pass rate ranged from 80% to 93% across the assessments for the Angoff method whilst by the Mean -1SD was 83% to 85%. We observed Cohen and Modified Cohen resulted in higher cut score which led to lower pass rates. For the Hofstee method, a similar pattern was seen and pass rates ranged from 75% to 86%. Table 3 portrays the fail rates for the studied standard setting methods. Angoff method generally resulted in a lower failure rate (below 20%). However, rest of the procedures showed wide variation and failure rates significantly higher (15% to 41%) than the Angoff except for the relative which yielded comparable failure rates.

The Angoff yielded the lower cut score which resulted in the consistent higher pass rates ranged from 83% to 95% in the middle/intermediate (Y3) clinical exams. On contrary Cohen and Modified Cohen yielded highest pass rate (91% to 98%) than the Angoff method. While for the Hofstee and relative method (Mean - 1SD) a similar pattern was observed in the pass rates (86% to 95%) and in determining the cut scores. We observed Hofstee, Cohen and Modified Cohen produced lower failure rates (approximately 8%) while the Angoff and relative method yielded comparable fails decisions that ranged from 5% to 15%. For the exit examination (Y5), both Angoff and relative methods produced comparable cut scores and pass rates that ranged from 81% to 86.8%. Cohen and Modified Cohen yielded similar cut scores but both produced approximately 100% pass rates. Though Hofstee method resulted in higher pass rates than the Angoff but slightly lower than the rest of methods. We noticed there were 15% more fails with the Angoff and relative methods than with the Cohen and Modified Cohen. A similar pattern of fail rates was observed for the rest of procedures as we found for the middle/intermediate clinical exams.

Table 3
Fail Rates (%) by each Standard Setting Method

	Exit/Final Clinical Exams			Early/Year 1 Clinical Exams			Middle/Intermediate Clinical Exams		
Methods	Y5 P1	Y5 P2	Y5 P3	Y1 P1	Y1 P2	Y1 P3	Y3 P1	Y3 P2	Y3 P3
Angoff	18.4	18.4	13.2	6.3	12.5	19.2	15.0	16.3	5.0
Cohen 65	5.3	2.6	1.3	37.5	35.0	41.8	6.3	8.8	1.3
Cohen 60	1.3	0.0	0.0	26.3	25.0	34.2	1.3	6.3	1.3
Mean-1SD	14.5	13.2	15.8	15.0	16.3	15.2	13.8	8.8	13.8
Modified Cohen 65	2.6	2.6	0.0	27.5	27.5	40.5	3.8	8.8	1.3
Modified Cohen 60	1.3	0.0	0.0	18.8	20.0	32.9	1.3	6.3	1.3
Hofstee	9.7	6.4	4.4	15.4	13.2	24.5	6.2%	7.3	5.6

Table 4

Pass Rates(%) by each Standard Setting Method

	Exit/Final Clinical Exams			Early/Year 1 Clinical Exams			Middle/Intermediate Clinical Exams		
Methods	Y5 P1	Y5 P2	Y5 P3	Y1 P1	Y1 P2	Y1 P3	Y3P1	Y3P2	Y3P3
Angoff	81.6	81.6	86.8	93.8	87.5	80.8	85.0	83.8	95.0
Cohen 65	94.7	97.4	98.7	62.5	65.0	58.2	93.8	91.3	98.8
Cohen 60	98.7	100	100.	73.8	75.0	65.8	98.8	93.8	98.8
Mean-1SD	85.5	86.8	84.2	85.0	83.8	84.8	86.3	91.3	86.3
Modified Cohen 65	97.4	97.4	100	72.5	72.5	59.5	96.3	91.3	98.8
Modified Cohen 60	98.7	100	100	81.2	80.0	67.1	98.7	93.7	98.7
Hofstee	90.3	93.6	95.6	84.6	86.8	75.5	93.8	92.7	94.4

Conclusions

In this study, we tested absolute method (modified Angoff), compromise method (Hofstee, Cohen and its modifications) and relative method (Mean, SD) to determine the cut scores for dental students to measure their competencies in the final, middle and early years of dentistry exams. We found meaningful differences in the passing scores of studied standard setting approaches as shown in Table 3 and Table 4 that indicates the proportion of passed and failed rates. In this study, there was a moderate agreement between the methods of standard setting in estimating the cut off scores across the assessments. The percentage agreement between the standard setting methods varied from 40% to 80% approximately. We found similar findings, which were reported in the literature for different format of assessments such as OSCE.

The modified Angoff procedure is a reliable and valid framework of determining cut off scores as it does not depend on the number of students. The accuracy of pass score based on how well examiners are trained to imagine the borderline candidate and define the characteristics of minimally competency but standardization can reduce the potential prejudice in their ratings. The Cohen and modified Cohen methods showed inappropriate fail rates (up to 40%) across the assessments in particular when assessments did not discriminate well among the students and had narrow score range. The plausible explanation could be non-normal score distribution, fewer number of students or students have achieved higher scores. We observed relative and Angoff showed comparable results while Hofstee and variation of Cohen procedure produced different results and resulted in higher cut scores which led to low pass rates that ranged from 58% to 80%. However, the major drawback is fixed percentage of students has to be failed and the difficulty of the exam does not taken into account. Similarly, estimates defined by examiners are beyond the expectations for the Hofstee method and the performance standards yielded are often too strict and may lead to higher number of failures. We can conclude for such health professional exams Angoff method is reliable in deciding the pass score irrespective of the years. We also found standard deviation of failure rates is lower as compared to the relative and compromised methods which shows that the modified Angoff gives a much more consistent failure and pass rates.

Wright (2016) suggested fail rates higher than 15% may considered as upper limit for high stakes exams such as health profession and entrance to certain degree programs. Our results suggested that safely we can conclude that Cohen method and its variations reported in the literature cannot be applied for standard setting in our context for the high stake assessments because these methods produced failure rates in excess of 15%. Further Cohen methods pass scores varied due to sample size and how well examinees performed in the upper quartile. We found in the literature that standard setting procedures produce different cuts scores but credibility, consistency and defensibility can be established by training of examiners and their ratings collected through systematic way (Impara & Plake, 1998). It is always challenging to answer which method is appropriate and what criteria we may apply to choose from available methods. There are number of factors such as purpose, context, type of test and level of stake play a role in the selection. One of the important check is to assess whether applied method have produced the desirable pass score, failure rates and expectations of the relevant stakeholders. Further investigation of the statistical characteristics of other methods is needed to ascertain or establish their merits and demerits and their applicability use in professional health related tests.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In: Educational measurement. Washington DC: American Council on Education; p. 514–515.

Cizek, G. J. (2012). An introduction to contemporary standard setting. In: Setting performance standards: foundations, methods, and innovations. 2nd ed. New York: Routledge.

Cohen-Schotanus, J. & Van der Vleuten, C. A. (2010). Standard setting method with the best performing students as point of reference: Practical and affordable. Medical Teacher, 32, 154-160.

De Gruijter, D. (1985) Compromise models for establishing examination standards. Journal of Educational Measurement, 22, 263-269.

Downing, S.M. & Yudkowsky, R. (2009). Assessment in health professions education. New York: Routledge.

Downing, S. M, Tekian, A. & Yudkowsky, R. (2006). Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. Teaching and Learning in Medicine, 18(1), 50-57.

Ebel, R. L. (1972). Essentials of educational measurement. Englewood Cliffs (NJ): Prentice Hall; p. 492-494.

George, S., Haque, M.S & Oyebode, F. (2006). Standard setting: comparison of two methods. BMC Medical Education, 6:46.

Humphrey-Murto, S. & MacFadyen, J.C. (2002). Standard Setting: A comparison of case author and modified borderline-group methods in a small-scale OSCE. Academic Medicine, 77, 134-137.

Hofstee, W, K, B. (1983). The case for compromise in educational selection and grading. In: On educational testing. Washington DC: Jossey Bass; p. 109-127.

Impara, J. C. & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. Journal of Educational Measurement. 35, 69-81.

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: theory and application. Educational Evaluation Policy Anal. 4, 461-476.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), Educational measurement (Third edition, pp. 485-514). Washington, DC: American Council on Education
and National Council on Measurement Education.

Kane, M. (1994). Validating the performance standards associated with passing scores. Review of Educational Research, 64, 425-461.

Kane, M., Crooks, T. & Cohen, A. (1999). Designing and Evaluating Standard-Setting Procedures for Licensure and Certification Tests. Advance Health Science Educational Theory Practice, 4**,** 195-207.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50, 1-73.

Kaufman, D. M., Mann, K.V., Muijtjens, A.M.M. & van der Vleuten, C.P.M. (2000). A comparison of standard setting procedures for an OSCE in undergraduate medical education. Acad Med. 75, 267-271.

Livingstone, S. A. & Zieky, M. J. (1982). Passing scores: a manual for setting standards of performance on educational and occupational tests. Princeton: Educational Testing Services.

McKinley, D.W. & Norcini, J.J. (2013). How to set standards on performance-based examinations: AMEE Guide No. 85. Medical Teacher. 36, 97-110.

Norcini, J. J. (2003). Setting standards on educational tests. Med Educ. 37, 464-469.

Nedelsky L. (1954). Absolute grading standards for objective tests. Educational Psychological Measurement. 14, 3-19.

Omer, A. E. & Karimeldin, M. A. S. (2015). Comparison of Two Standard Setting Methods in a Medical Students MCQs Exam in Internal Medicine. American Journal of Medicine and Medical Sciences, 164-167.

Taylor, C. A. (2011). Development of a modified Cohen method of standard setting. Medical Teacher, 33, 678 - 682.

The General Dental Council (2012). Standards for education. London

Wass, V., van der Vleuten, C.P.M., Shatzer, J. & Jones, R. (2001). Assessment of clinical competence. Lancet. 357, 945-949.

Wright, J. D. (2016). The Application of Simple, Practical and Affordable Standard Setting Methods in Small Groups of Students. Medical Education Publisher. 5, 106.

Wood, T.J., Humphrey-Murto, S.M. & Norman, G.R. (2006). 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. 11,115-122.

Zieky, M. J. (2001). So much has changed: How the setting of cut-scores has evolved since the 1980s. In G. J. Cizek (Ed.), Setting performance standards (pp. 19-52). Mahwah, NJ: Lawrence Erlbaum.

....✳✳✳....

Citation of this Article:

Khalid, M. N., Shafiq, F. & Ahmad, S. (2021). A Comparison of Standard Setting Methods for Setting Cut-Scores for Asessments with Constructed Response Questions**.** *Pakistan Journal of Educational Research and Evaluation, 9*(2), 74-85.