Implementing the Angoff method of standard setting using postgraduate students: Practical and affordable in resource-limited settings

A G Mubuuke, BSc, MSc, MPhil, PhD Fellow; **C Mwesigwa**, BDS, MSc; **S Kiguli**, MB ChB, MMed, MHPE

College of Health Sciences, Makerere University, Kampala, Uganda

Corresponding author: *A G Mubuuke (gmubuuke@gmail.com)*

Background. Cut scores for students' assessments have always been arbitrarily determined in many institutions. Some institutions have adopted reliable methods of determining cut scores, such as the Angoff method. However, use of this method requires many experts, making it difficult to implement in resource-limited settings. The possibility of involving postgraduate students in implementing the Angoff method of setting cut scores could be the solution to this problem.

Objectives. To explore the knowledge and practices of faculty regarding standard setting and the feasibility of using postgraduate students when implementing the Angoff method.

Methods. This was an exploratory operations research study in which data were collected during focus group discussions. Students were trained to use the Angoff method, i.e. a previous examination, in which the pass mark was 50%, was used to evaluate the method.

Results. Initial findings showed that faculty in the consortia of schools did not know what standard setting and the Angoff method entailed and had never used this approach. The postgraduate students involved in implementing the Angoff method of setting cut scores were excited and interested in engaging in the exercise; the pass mark they arrived at was 61.21%.

Conclusion. The study demonstrated that it is feasible to use the Angoff method of determining pass marks, even in resource-limited settings. This can be made possible by involving postgraduate students in the absence of enough faculty experts.

Afr J Health Professions Educ 2017;9(4):171-175. DOI:10.7196/AJHPE.2017.v9i4.631

During curriculum development, teachers adopt various criteria to assess the students' level of competence. These are put in practice during tests and examinations. Ideally, examiners need an educational method to determine cut scores to distinguish non-competent from competent students. The practice of determining cut scores is called standard setting.[1,2] A cut score is a point on a scale that separates one performance standard from another. The traditional arbitrary methods used to define cut scores, such as responding correctly to 50% of the test items, cannot provide robust and valid evidence to judge student performance. Therefore, the use of these methods may be difficult to justify. Consequently, there is a need to set cut scores using methods that are robust, valid, and provide a fair judgement of student performance. Although no method has been identified as the benchmark for setting cut scores, the use of scientific methods with a systematic approach provides a balanced judgement of student performance.[3-5]

There are two broad categories of standard setting: (*i*) the criterion or absolute method, where setting a cut score is independent of test results;[6-9] and (*ii*) the norm-referenced or relative method, where cut scores are set depending on test results.[6-9] Norm-referenced methods are generally used to rank students, while criterion-referenced methods are used to judge student performance against a set benchmark.[7-9] The criterion-referenced methods for setting cut scores in health professions education usually involve a number of subject experts making judgements about test items and proposing a final cut score; this is labour intensive, costly and subjective. The current study focused on a feasible way of using the Angoff method of setting cut scores in resource-limited settings with few experts.

The original Angoff and modified Angoff methods have been widely used in setting cut scores.[8] The original method requires a panel of subject experts to determine the probability of a minimally competent student answering a test item correctly. It requires each expert to estimate the probability of each test question. The final cut score becomes the average of the sums of different probabilities from all experts.[10] In the original Angoff method, experts determine the probabilities, i.e. they can select any probability ranging from 0 to 1 (0.90, 0.44, 0.56, etc.). The modified Angoff method restricts the probabilities to eight choices (0.2, 0.4, 0.5, 0.6, 0.75, 0.90, 0.95, 'do not know').[7, 11]

The Angoff method of setting cut scores is resource intensive and requires many well-qualified experts in the test domain. In many institutions, there are not enough qualified experts to form a reliable panel in any one particular field. The few available have to divide their time between many tasks other than student assessment.[12,13] One needs to find a way of effectively using the available resources to implement the Angoff method in a resource-limited context. This study had two purposes: (*i*) to explore the knowledge and practices of faculty about standard setting and the use of the Angoff method; and (*ii*) to explore the feasibility of using postgraduate students as panel members when implementing the Angoff method of standard setting.

Methods

Study setting

The study took place in Uganda and involved faculty from five medical schools: Makerere University College of Health Sciences (MaKCHS), Kampala; Mbarara University of Science and Technology; Gulu University; Busitema University; and Kampala International University. Under the auspices of the Medical Education Partnership Initiative (MEPI), the five Ugandan medical schools formed a consortium – the Medical Education for

Equitable Services for All Ugandans (MESAU) – to have one unified voice aimed at improving the training of health professionals in the country. This consortium developed common competencies and suggested the adoption of common assessment practices.

Study design

This was a hands-on research study in which knowledge and practices of lecturers regarding standard setting and the Angoff method were initially explored across the MESAU schools during focus group discussions. After conducting a baseline exploration of lecturers' knowledge and practices of standard setting, we investigated the feasibility of using postgraduate students as part of the panel of experts to set cut scores for undergraduate students, employing the original Angoff method. This was done as a pilot study in the radiology department of one of the MESAU schools. Six postgraduate students in this department and two faculty members were recruited through convenience sampling to participate in the scoring of examination questions.

Before the scoring exercise, three short training sessions, one per day, were organised for the relevant students and faculty. Each training session lasted ~25 minutes and focused on the meaning of standard, advantages and using the original Angoff method to set cut scores. Scheduling of the training sessions into three short sessions allowed the postgraduate students time for other learning activities. In the last session, the 6 postgraduate students and 2 faculty members were briefed about the exercise and possible issues were clarified.

The following day, the relevant postgraduate students and faculty members were invited to form a panel of experts (also referred to as judges), who would score the test questions and provide a final pass mark. A previously written test for undergraduate students was used for the exercise. This test had 30 questions; a student had to circle one single-best correct option. To avoid bias, the correct answer was not shown to the judges. The key guiding question for the panel during the exercise was: What is the percentage chance of a borderline student answering this question correctly? The researchers carefully formulated the question using simple language. They avoided educational terminologies because the intended audience comprised non-educational experts. Therefore, the researchers further defined a borderline student as one who spends a minimum of time studying, is good enough to pass the examination and often finds it difficult to score above the pass mark.

Each judge was then requested to note down any percentage chance for each test question for all 30 questions. After the initial round of scoring, the judges discussed the scores among themselves. The facilitator also afforded each group the opportunity to express their opinions. Subsequently, a second round of scoring was done, the various average scores from the 8 judges were compiled, and a final cut score for the test was determined.

After setting the cut score for the test, the 6 postgraduate students who participated in the exercise were invited to participate in a 30-minute focus group discussion the following day to share their experiences. One key assumption was that a postgraduate student in radiology had the required competency to determine whether a borderline undergraduate medical student can answer a given radiology question correctly.

Data collection and analysis

Focus group discussion was the primary method of collecting data in each MESAU school. Two focus group discussions, which included the lecturers, were conducted in each school, giving a total of 10 focus group discussions conducted across the 5 MESAU schools. Each focus group comprised 6 participants. The researchers audio recorded and later transcribed the responses from these discussions. Two of the researchers then read through the data. Thematic analysis was used,[14] and the researchers analysed the data manually. During this process, raw data were read, and through a series of iterative and inductive open and axial coding, codes and themes were developed manually.[15]

Quality assurance

The researchers stored the data electronically and secured these with a password. Participants were invited to validate the emerging themes to ensure credibility of the data. Additionally, researcher bias was minimised by the researchers, avoiding all preconceived ideas or experiences on the subject being investigated and practising reflexivity and bracketing throughout the research process.

Ethical considerations

Participants provided written informed consent. They were not identified by name and their responses were kept anonymous and confidential. Permission to conduct this study was granted by the Research and Ethics Committee, School of Health Sciences, MaKCHS (ref. no. 2014-045).

Results

The lecturers generally had limited knowledge of standard setting and mostly did not practise it. One major theme arose from the analysis, with key representative responses, as indicated below.

Knowledge and practices of lecturers regarding standard setting

The lecturers who participated in the focus group discussions lacked knowledge of standard setting in assessment, almost all of them agreeing that they did not know what it means. A few had heard about the concept, but did not know what it entailed. Some typical responses are given below:

> 'I have not heard about standard setting and cannot tell what it exactly means.'
>
> 'I have heard about standard setting from a few seminars and workshops I have attended – that it involves setting pass marks. However, I feel am not competent enough to explain what it is.'
>
> 'I am not an education expert and therefore I cannot commit myself to offer an explanation as to what standard setting means.'

From the responses listed above and many more that echoed a similar interpretation, it is clear that lecturers involved in student assessment lacked knowledge of standard setting. Moreover, the lecturers had never practised standard setting in their institutions during assessment:

> 'Why should I practise what I do not know?'
>
> 'We cannot practise standard setting unless someone teaches us what it is and how it should be done.'
>
> 'Although I have a little knowledge about standard setting, I have never practised it myself.'

From the responses it was therefore clear that the faculty members who participated did not know what standard setting is, and had never practised it. Additionally, none of the lecturers had ever heard about the Angoff method of setting a cut score:

> 'We have never heard about that terminology and do not know what it means.'

Using postgraduate students, it was observed that the final cut score determined from the scoring was 61.21%. Table 1 illustrates how each test question was scored by each judge, the various averages of the raters, as

Table 1. Scores (%) from each judge and final cut-score

	Judge and score, %								
Question	1	2	3	4	5	6	7	8	Cut score, mean (SD)
1	50	55	61	60	58	50	49	55	54.75 (4.71)
2	60	60	55	60	58	57	60	55	58.12 (2.23)
3	54	58	60	50	55	55	50	50	54.00 (3.82)
4	50	50	52	53	58	60	60	55	54.75 (4.17)
5	50	45	50	55	55	57	60	60	54.00 (5.29)
6	50	50	50	56	53	54	60	60	54.12 (4.22)
7	55	60	60	60	55	50	50	50	55.00 (4.63)
8	54	56	55	55	60	60	60	55	56.88 (2.64)
9	55	55	50	50	60	65	65	70	58.75 (7.44)
10	52	53	55	55	60	60	50	56	55.13 (3.56)
11	65	60	60	70	65	65	60	60	63.13 (4.58)
12	50	50	48	50	52	50	50	52	50.25 (1.28)
13	65	58	60	65	60	55	60	60	60.38 (3.34)
14	60	60	54	55	53	60	65	70	59.63 (5.78)
15	90	80	80	75	85	80	80	75	80.63 (4.96)
16	70	75	70	65	70	78	80	65	71.63 (6.58)
17	60	57	65	60	70	65	60	55	61.50 (4.87)
18	85	80	78	80	90	75	85	80	81.63 (4.75)
19	100	85	88	95	90	85	80	80	87.88 (7.00)
20	56	50	49	55	50	50	55	50	51.88 (2.90)
21	60	58	65	60	60	50	55	55	57.88 (4.52)
22	70	67	65	60	75	65	65	70	67.13 (4.52)
23	56	55	60	50	50	55	52	60	54.75 (3.96)
24	70	65	60	75	60	60	58	60	63.50 (6.05)
25	85	80	80	75	80	85	80	78	80.38 (3.34)
26	55	60	53	50	51	50	55	50	53.00 (3.55)
27	65	60	58	55	60	64	60	60	60.25 (3.15)
28	70	65	75	60	60	70	65	60	65.63 (5.63)
29	55	58	70	54	56	60	60	55	58.50 (5.18)
30	55	58	50	49	45	48	51	55	51.38 (4.31)
Final average cut score for minimum competency									
	62.40	60.77	61.20	60.40	61.80	61.27	61.33	60.53	61.21 (9.88)

SD = standard deviation.

well as the standard deviations (SDs) from the mean scores. From the SDs, it can be observed that across the test items, there was generally no large dispersion of scores from the mean. Also, the final cut score fell within the mean cut score of each judge for the 30 questions. The entire exercise of setting the cut score lasted 90 minutes. Having participated in the exercise, a focus group discussion was conducted with the postgraduate students to explore their experiences. The findings are presented below.

Experiences of postgraduate students after the scoring exercise

The focus group discussion conducted with postgraduate students after the standard-setting exercise revealed interesting and encouraging experiences. All postgraduate students who participated expressed excitement about becoming involved, as can be observed in the following responses:

'This was a whole new experience to me. It was indeed interesting for me to get involved in determining other students' pass marks. I wish ours were determined like this before.'

'This is the best way to go and I thank our teachers for getting us this opportunity. I feel that this system is fair to students and will be welcomed if implemented fully.'

'We enjoyed the whole exercise. This method of determining pass marks where people follow a systematic process is not only fair, but also acceptable. Just saying that the pass mark is 50% does not make sense.'

From such responses, it appears that the graduate students enjoyed the exercise and supported setting a pass mark using the relevant steps.

Although they generally accepted the method, the graduate students expressed some concerns:

> 'This is very good. However, I have seen that one needs several lecturers to do it.'
>
> 'The exercise of setting the pass mark required some time. In my opinion, time considerations need to be put in place before carrying out the exercise, like setting exams early enough and determining the pass mark before students sit for the exam.'
>
> 'Availability of time is the most crucial thing here. Do lecturers have enough time to carry out this exercise?'

The abovementioned responses single out the factor of time, which should be considered when planning implementation of this exercise. However, the graduate students had a solution to mitigate this:

> 'Like we sometimes do participate in teaching of undergraduates, we can also participate in determining a pass mark alongside our lecturers. If this exercise is carried out early enough before exams commence, the time factor can be fairly addressed.'
>
> 'We can dedicate some time on our timetables to participate in determining a pass mark for our undergraduate fellows. At least, if it is time tabled, there should be no problem. Indeed, reserving a little time to participate also refreshes our memories of what we learned earlier.'

From the responses, it appears that graduate students were eager to participate and allow some time for this exercise.

Discussion

This study explored lecturers' knowledge and practices of standard setting across MESAU schools and the possibility of using postgraduate students in the standard-setting process.

Knowledge and practices of lecturers regarding standard setting

Findings of the current study illustrated that the lecturers lacked adequate knowledge of standard setting, specifically of the Angoff method, which they did not practise before. While this was a significant observation, it may not be surprising. Many lecturers in these medical schools lack formal training in medical education and are not very conversant with issues of standard setting. The majority are recruited into teaching owing to excellent grades in their professional disciplines, which do not involve educational issues. This probably explains the observation that they lacked knowledge about standard setting.

Our study also points to a lack of adequate faculty-development programmes in standard setting in these MESAU institutions. Many lecturers in medical schools lack educational knowledge and skills; this is not unique to the MESAU schools, but has been widely reported elsewhere.[7] Many institutions have taken on the initiative to design and implement faculty-development programmes, targeting specific faculty needs to improve teaching, learning and assessment.[13]

Feasibility of employing postgraduate students in the standard-setting process

The study also explored the feasibility of implementing the original Angoff method using postgraduate students. Findings indicated that they fully participated in and were very excited about the exercise. One would have expected these students to complain about the additional workload alongside their usual learning activities. It is, however, not clear why postgraduate students were excited and found the exercise interesting. One can argue that it probably benefited them educationally, as it allowed them to revise and refresh their memories with regard to previous learning material. One can also argue that as their own cut scores were predetermined when they were students, they were eager to participate in the process of determining cut scores for their colleagues.

Moreover, it appears as if the standard-setting process provided what could be deemed a credible cut score for the test, despite the participation of postgraduate students as judges. The final cut score for the test used in this study was 61.21%, whereas a cut score of 50% had previously been used for this test. The cut score of 61.21%, as determined by the panel of judges, seems a fair, valid and reliable representation of the difficulty of the test compared with the 50% score. This can be supported by previous records, which show that the lowest-scoring student in this particular test achieved 63%, which is above our cut score of 61.21%, determined by the Angoff method. This vindicates our exercise and suggests that the Angoff method had some degree of reliability and credibility. This observation is in agreement with findings from Verhoeven *et al.*,[1] who reported that using recent graduates as judges when implementing the Angoff method can be credible and reliable.

One could argue that postgraduate students are not subject experts. However, all such students have studied the undergraduate curriculum and should have the minimum competency to offer an opinion regarding the probability of an average undergraduate student answering a question correctly.

The advantage of the Angoff method is that judges can initially score the questions and then discuss their scores before continuing with another round of scoring. With the exercise taking place in the presence of two faculty members, the discussion most probably offered valuable insights, which encouraged the participating postgraduate students to reflect on and think carefully about their initial scores and the test items before the second round of scoring.

To tap into the advantages of the Angoff method while simultaneously not overburdening the few available academic staff, this study proposes involving postgraduate students in various departments to become part of the panels, together with some faculty members, as a way of implementing the Angoff method in the context of limited human resources. However, the postgraduate students need to be trained alongside faculty so that they know what is expected of them.

The issue of time, as observed from the responses, should not be overlooked, as the exercise can appear as an additional workload to the already busy students. It is suggested that faculty need to take into consideration postgraduate students' time. It was feasible to divide the training into three short sessions of 25 minutes per day for 3 days, instead of a 2-hour session for 1 day. The suggestion from the participating students that examinations be set early and the exercise be time tabled is another way of addressing the time factor. Furthermore, postgraduate students could receive an assessment mark for participating in this exercise as a way of motivating them. Without proper scheduling of time, taking into consideration postgraduate students' learning periods, their involvement is not likely to succeed.

From the literature, it appears that there are no studies exploring the possibility of postgraduate students as judges when setting cut scores, using the original Angoff method, in the event of limited academic staff. Although Verhoeven *et al.*[1] studied this aspect using recent graduates on progress tests, they employed the modified Angoff method and provided

correct answers to the judges before the scoring exercise, an observation that arguably creates bias. By providing the correct answer, the mind of the judge is influenced and a seemingly difficult question might be viewed as easy, and vice versa, which creates some degree of bias.

We decided not to provide answers to the judges to avoid such a scenario. Additionally, the modified Angoff method that Verhoeven *et al.*[1] used also restricts judges to specific scores.[7] The disadvantage is that judges are limited to the use of predetermined scores, which can be viewed as a way of influencing their decision. We left the scoring open, so that the judges could carefully consider the question and provide an appropriate score from a very wide range of possible scores. Therefore, the contribution of this study is worth noting and building on. Our approach of training the judges before the exercise most probably eliminated all uncertainties in the minds of the judges; therefore, it was clear what was expected of them. This eliminated issues of providing correct answers to the judges, which could lead to bias. Nonetheless, findings from our study generally concur with those of Verhoeven *et al.*[1] and further illustrate that postgraduate students can be judges when using the Angoff method. In our study, short training sessions for the student judges possibly eliminated the requirement of providing correct answers to them when scoring. Simple, short training sessions, e.g. half a day, are specifically encouraged. Because of these observations, we suggest using postgraduate students as part of the panel that determines cut scores for undergraduate students in situations where there are not enough subject experts to form such a panel in a resource-limited setting.

Study limitations

We used postgraduate students in only one department, a major limitation of the study. It is difficult to recommend a major roll-out using data from only one department. We therefore suggest that such an exercise be tried and evaluated in other departments, and incremental implementation be carried out rather than a major roll-out at the MESAU schools and other schools. However, the information gathered provides a foundation on which this exercise can be applied elsewhere and findings compared.

Further research

A major focus of this study was addressing the human resource gap when using the Angoff method; it did not specifically focus on how time can be used optimally when involving postgraduate students. This provides a direction for future research.

Conclusion

Our study has demonstrated that postgraduate students can be efficiently used as a cost-effective measure to address the human resource gap when employing the Angoff method of setting cut scores. There is also a need for faculty-development programmes in assessment and standard setting, so that faculty can have a basic knowledge of what these programmes entail. In this manner, the advantages of introducing innovations, such as standard setting, are most likely to be reasonably well accepted instead of being completely rejected.

Acknowledgements. We acknowledge support from the Medical Education for Equitable Services for All Ugandans-Medical Education Partnership Initiative (MESAU-MEPI). The content is solely the responsibility of the authors and does not necessarily represent the official views of the Fogarty International Center or the National Institutes of Health.

Author contributions. AGM: conceived the idea, drafted the protocols for ethical reviews, participated in designing the study tools and in data collection and analysis, and wrote the initial draft; CM: refined the idea, participated in designing the study tools and in data collection, and reviewed the initial draft; SK: guided the team during the process and reviewed the final draft.

Funding. This study was funded by the MESAU-MEPI Programmatic Award (ref. no. 1R24TW008886) from the Fogarty International Center.

Conflicts of interest. None.

1. Verhoeven BH, van der Steeg AFW, Scherpbier AJJA, Muijtjens AMM, Verwijnen GM, van der Vleuten CPM. Reliability and credibility of an Angoff standard setting procedure in progress testing using recent graduates as judges. Med Educ 1999;33(11):832-837. https://doi.org/10.1046/j.1365-2923.1999.00487.x
2. Muijtjens AMM, Schuwirth LWT, Cohen-Schotanus J, Thoben AJNM, van der Vleuten CPM. Benchmarking by cross-institutional comparison of student achievement in a progress test. Med Educ 2008;42(1):82-88. https://doi.org/10.1111/j.1365-2923.2007.02896.x
3. Friedman B-D M. AMEE Guide No. 18: Standard setting in student assessment. Med Teach 2000;22(2):120-130. https://doi.org/10.1080/01421590078526
4. Taylor CA. Development of a modified Cohen method of standard setting. Med Teach 2011;33(12):e678-e682. https://doi.org/10.3109/0142159X.2011.611192
5. McHarg J, Bradley P, Chamberlain S, Ricketts C, Searle J, McLachlan J. Assessment of progress tests. Med Educ 2005;39(2):221-227. https://doi.org/10.1111/j.1365-2929.2004.02060.x
6. Norcini J, Guille R. Combining tests and setting standards. In: Norman GR, van der Vleuten CPM, Newble DI, eds. International Handbook of Research in Medical Education. Dordrecht, The Netherlands: Kluwer Academic Publishers, 2002.
7. Norcini J. Setting standards on educational tests. Med Educ 2003;37(5):464-469. https://doi.org/10.1046/j.1365-2923.2003.01495.x
8. Bandaranayake RC. Setting and maintaining standards in multiple choice examinations: AMEE Guide No. 37. Med Teach 2008;30(9):836-845. https://doi.org/10.1080/01421590802402247
9. George S, Sayeed Haque M, Oyebode F. Standard setting: Comparison of two methods. BMC Med Educ 2006;6:46. https://doi.org/10.1186/1472-6920-6-46
10. Angoff WH. Scales, norms, and equivalent scores. In: Thorndike RL, ed. Educational Measurement. 2nd ed. Washington, DC: American Council on Education, 1971:508-600.
11. Gagnon R, Charlin B, Coletti M, Sauvé E, van der Vleuten CPM. Assessment in the context of uncertainty: How many members are needed on the panel of reference of a script concordance test? Med Educ 2005;39(3):284-291. https://doi.org/10.1080/10401334.2010.488197
12. Verhoeven BH, Verwijnen GM, Muijtjens AMM, Scherpbier AJJA, van der Vleuten CPM. Panel expertise for an Angoff standard setting procedure in progress testing: Item writers compared to recently graduated students. Med Educ 2002;36(9):860-867. https://doi.org/10.1046/j.1365-2923.2002.01301.x
13. Prince KJAH, Scherpbier AJAA, van Mameren H, Drukker J, van der Vleuten CPM. Do students have sufficient knowledge of clinical anatomy? Med Educ 2005;39(3):326-332. https://doi.org/10.1111/j.1365-2929.2005.02096.x
14. Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Med Res Methodol 2008;8:45. https://doi.org/10.1186/1471-2288-8-45
15. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. https://doi.org/10.1191/1478088706qp063oa

Accepted 13 June 2017.