The Angoff method in the written exam of the College of Intensive Care Medicine of Australia and New Zealand: setting a new standard

Christian Karcher

In any test, a cut-point or standard needs to be established to discriminate between acceptable and unacceptable examinee performance. In the case of the Fellowship exam of the College of Intensive Care Medicine of Australia and New Zealand (CICM), the written component functions as a filter to exclude candidates who are deemed not yet ready for the clinical component examination.

At present, the written part of the CICM Fellowship examination, consisting of 30 short-answer questions (SAQ), has a pass mark of 50%. Each question is worth ten points (300 points in total), with all questions carrying equal marks. An arbitrarily chosen standard of 50% means that all written examinations have to then be designed with this target in mind.[1]

It has long been a criticism of the written examination that the degree of difficulty varies from exam sitting to exam sitting, and it has been suggested that this was the main reason for variable overall success rates for the candidates. Between October 2012 and October 2017, the mean pass rate of the written exam was 66%, ranging from 43% (May 2014) to 80% (October 2017).[2] Assuming that the cohorts of candidates presenting to those exams were comparable with regards to their knowledge and that the standard of marking was constant, this variation in pass rates may indeed indicate that the written examinations were not equally difficult.

The Board of the CICM recently endorsed a major change to the marking of the written component of the fellowship exam.[3] With the first sitting of the 2019 exam, the College will introduce the Angoff method to determine the pass/fail mark for the written part.

The Angoff method is the oldest absolute standard method and has been published widely. An absolute standard method means that the number of successful candidates is not predetermined; hence, theoretically, all or none could pass the test based on whether they meet or not the acceptable standard. In contrast, in a relative standard method, a set proportion of candidates will pass the test (eg, the top 60% of the cohort).

When applying the Angoff method, a panel of subject-matter experts (SMEs) initially describes a minimally competent candidate (MCC). The MCC is, by definition, a candidate that would achieve the minimum standard to pass. Such a candidate would have a 50/50 chance of passing the item or question.[4,5] In colloquial terms, this candidate would be "just good enough".

The SMEs would then determine the probability of an MCC answering the examination item correctly, which is commonly done using a *P* value between 0 and 1. Alternatively, for every item, the SMEs can determine whether the MCC will answer it correctly or not.[6]

While the literature on Angoff standards for multiple-choice questions is quite extensive, there is a surprising sparsity of literature specific to its use with regards to short answer questions.[7]

The Australasian specialist Colleges for emergency medicine (ACEM), general practice (RACGP), obstetrics and gynaecology (RANZCOG), surgery (RACS), medicine (RACP) and rural medicine (ACCRM) and the European Society for Intensive Care Medicine all use the Angoff method.[3,8-10]

For a SAQ exam, determining pass/fail is obviously not suitable, since the result for SAQ is not dichotomous (eg, "true or false"). The SMEs are therefore required to determine what percentage of the maximum score an MCC would achieve. For this to be a valid method, the SME must have a detailed marking template for each question with key answer components being highlighted and associated with marks. The questions do not necessarily have to carry the same maximum mark. In fact, weighting of topics across the exam can be achieved by assigning a higher or lower maximum mark to questions depending on the importance of the topic.

It has been described in the literature that teachers have difficulties predicting their students' performance, and judges' ratings may vary greatly.[11-13] The authors suggest that this may be related to the judges' familiarity with the standard setting process and their knowledge of the curriculum and the learners, rather than to real differences in perceived item difficulty.[14,15]

It follows then that CICM examiners face the same challenge when it comes to predicting the performance of intensive care trainees. The SMEs' ability to accurately predict a candidate's performance may be influenced by a number of issues. First, the examiner might be biased towards a specific trainee who they are familiar with and, hence, have prior knowledge of their strengths and weaknesses. Therefore, this preconceived "mental model" of an MCC may lack external validity. Second, the CICM examiner may be biased by their own knowledge; that is, perceiving a question as difficult for themselves may lead to underestimation of the MCC's score. Conversely, a question that is perceived as easy by the examiner (eg, from a topic of their particular interest) may lead to overestimation of the MCC's mark. Finally, the historic 50% pass mark may still be so deeply ingrained that the SME may gravitate towards it.

Once the SMEs have determined the pass mark for an item, all marks for that item are averaged. A second rating round aiming at reduction of inter-rater differences and increased consensus can eliminate some, but not all, of those three effects described above.

Once a pass mark for each item (question) has been agreed upon, the overall pass mark of the exam is finally calculated as the sum of all 30 agreed pass marks.

Despite its challenges, the Angoff method is relatively easy to implement. An appropriately large number of SMEs is required to maintain validity. An adequate selection of judges is crucial for the process of standard setting. The judges' attributes determine the values they apply to defining the minimally competent candidate. A recommended minimum number of SMEs is six to eight.[16] Especially in high-stakes tests, such as a final fellowship examination, reliability is determined not only by the number of SMEs but also by their backgrounds and perspectives. The CICM will have to ensure that the group of SMEs is carefully selected, with a balanced distribution of experience, age, gender, ethnicity as well as location and focus of their primary workplace.

At CICM, seven to ten fellowship examiners have been chosen from the examiner's group to act as SMEs for the Angoff method (Associate Professor Jeremy Cohen, Chair CICM Second Part Exam Committee, personal communication, August 2018). The fact that the examiners, who are all practising intensive care specialists, write, test and review the written questions ensures a high level of relevance to the examination.

Among the judges, a broad agreement on the validity of the process is required to avoid less serious effort and lapses during the rating process.[15]

Since the standard setting process is not an anonymous one, there is a risk of the individual SME being biased either by the group or the behaviour or "strong" individuals.[17,18] The latter is more problematic. A very assertive member of the SME group may influence not just one other SME but many others, resulting in the group consensus shifting towards the outlier.

Conversely, a "weak" outlier could converge towards the central tendency of the group through conformity pressure, thereby increasing the level of consensus.

The total CICM Fellowship examination marks are currently calculated by adding the marks of the written component after weighting (maximum, 30 marks; pass mark, 15) and the clinical component, consisting of both the cross-table vivas (maximum, 40 marks; pass mark, 20) and the clinical hot cases (maximum mark, 30; pass mark, 15). The proposed changes to the written exam will have an impact on the overall exam mark. Since the pass mark will vary between cohorts, a candidate can now pass the written exam with less than 50% (less than 15 total marks). Hence, the written mark can no longer contribute towards the total exam mark for comparison with other cohorts of candidates; however, this will not affect the conduct of the exam or the awarding of the College medal for the highest score.

The previously used, arbitrarily chosen, fixed cut-off mark was considered non-defensible and unfair. Failure and success were not only influenced by the candidates' knowledge but also by the variation in the degree of difficulty of the exam.

From a candidate's perspective, one other advantage of the Angoff method over other standard setting techniques is that the standard can be set before the candidates sit the examination, thus, decreasing the time between the written examination and the release of results. As mentioned previously, passing the written examination is a prerequisite for the clinical examination, so the time between receiving the written examination results and sitting the clinical examination can thereby be maximised. Not only is this crucial for the logistics of the clinical component but also for the preparation time that successful candidates have between the two examination components.

To change the pass mark in the CICM written Fellowship exam from a 50% standard to one established by using the Angoff method appears to be a very reasonable decision, as it provides a reliable, validated, fair and defensible standard.

The workload to establish the standard for each exam paper seems to be manageable for the group of examiners and the administration.

The College will have to implement an ongoing, rigorous review process once the new standard method is introduced.

Competing interests

None declared.

Author details

Christian Karcher[1,2]

1 Department of Intensive Care, The Royal Melbourne Hospital, Melbourne, VIC, Australia.

2 University of Melbourne, Melbourne, VIC, Australia.

***Correspondence*:** Christian.Karcher@mh.org.au

References

1 Tekian A, Norcini J. Overcome the 60% passing score and improve the quality of assessment. *GMS Z Med Ausbild* 2015; 32): Doc43.

2 College of Intensive Care Medicine of Australia and New Zealand. Second Part Examination: exam reports [website]. CICM; 2018. https://www.cicm.org.au/Trainees/Assessments-and-Examinations/Second-Part-Exam#PreviousExamReports (viewed May 2018).

3 College of Intensive Care Medicine of Australia and New Zealand. Introduction of Angoff marking system to the Second Part Examination [website]. CICM; 2018. https://www.cicm.org.au/News-Summary/Introduction-of-Angoff-Marking-System-to-the-Secon (viewed June 2018).

4 Norcini JJ. Setting standards on educational tests. *Med Educ* 2003; 37: 464-9.

5 Carlson J, Tomkowiak J, Stilp C. Using the Angoff Method to set defensible cutoff scores for standardized patient performance evaluations in PA education. *J Physician Assist Educ* 2009; 20: 15-23.

6 Impara JC, Plake BS. Standard setting: an alternative approach. *J Educ Meas* 1997; 34: 353-66.

7 Reid KJ, Dodds AE, Fink MA. Setting short–answer question standards using borderline regression. *Med Educ* 2015; 49: 520-1.

8 McCall L, Foley B. The new ACEM Fellowship examination: lessons learnt. *Emerg Med Australas* 2016; 28: 232-5.

9 Chan SC, Mohd Amin S, Lee TW. Implementing standard setting into the conjoint MAFP/FRACGP part 1 examination — process and issues. *Malays Fam Physician* 2016; 11): 2-8.

10 Serpell JW. Evolution of the OSCA-OSCE-clinical examination of the Royal Australasian College of Surgeons. *ANZ J Surg* 2009; 79: 161-8.

11 Impara JC, Plake BS. Teachers' ability to estimate item difficulty: a test of the assumptions in the Angoff standard setting method. *J Educ Meas*1998; 35: 69-81.

12 Fitzpatrick AR. Social influences in standard setting: the effects of social interaction on group judgments. *Rev Educ Res* 1989; 59: 315-28.

13 Goodwin LD. Determining cut-off scores. *Res Nurs Health* 1996; 19: 249-56.

14 Mills CN, Melican GJ, Ahluwalia NT. Defining minimal competence. *Educ Meas* 1991; 10: 7-10.

15 Plake BS, Melican GJ, Mills CN. Factors influencing intrajudge consistency during standard-setting. *Educ Meas*1991; 10: 15-6.

16 Brennan RL, Lockwood RE. A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. *Appl Psychol Meas* 1980; 4: 219-40.

17 Barsade SG. The ripple effect: emotional contagion and its influence on group behavior. *Adm Sci Q* 2002; 47: 644-75.

18 Cartwright D, Lippitt R. Group dynamics and the individual. *Int J Group Psychother* 1957; 7: 86-102.