

Submitted: 14 March 2024
Accepted: 13 November 2024
Published online: 1 April, TAPS 2025, 10(2), 91-93
<https://doi.org/10.29060/TAPS.2025-10-2/II3264>

Enhancing standard setting: A judge's guide for the Angoff method in assessing borderline students

Han Ting Jillian Yeo & Dujeeepa D. Samarasekera

Centre for Medical Education (CenMED), Yong Loo Lin School of Medicine, National University of Singapore, Singapore

I. INTRODUCTION

Assessment is an important component of training in ensuring that graduating students are competent to provide safe and effective medical care to patients. Typically, the passing score is set as a fixed mark, but this approach does not account for the varying difficulty of exams. As a result, students who have achieved the required level of competence might fail if the exam items are particularly challenging (false negative), while students who have not attained the necessary competence might pass if the items are unusually easy (false positive). Hence, deciding on the right pass mark is important for each assessment. To mitigate this issue, criterion referenced standard setting was adopted in medical education (Norcini, 2003). It determines the minimum competence level expected of a candidate and whether a candidate would pass or fail the assessments (Norcini, 2003). The Angoff method is one of the more commonly used standard setting techniques. It is an examinee centred method and requires a panel of judges to estimate the probability that a borderline candidate would get the item correct.

Literature have questioned the reliability of the Angoff method. Variations in pass mark have been reported when the different panels of judges were engaged (Tavakol & Dennick, 2017; Taylor et al., 2017). Judges reportedly faced challenges in visualising and defining the knowledge and skills required of borderline students and hence have difficulty estimating the probability that a borderline student would answer an item correctly (Tavakol & Dennick, 2017). A study by Yeates et al.

(2019) also reported the complexity judges faced in the standard setting process due to interaction between the environment, individual judgments, and interaction between the judges. Such variations in pass marks might lead to unfairness to students who were meant to pass but did not due to a higher pass mark. It is of a greater concern to patient safety if students who were meant to fail passed the examination due to a lowered pass mark. To assist the judges, a guide was developed to set standards for medical and health professions examinations using a probability estimate.

II. DEVELOPING A GUIDE

Judges were to rate each item based on three criteria: relevance, frequency, and difficulty. The guide focused on these areas to assist the judges in their evaluations. The relevance of an item was rated on a 5-point scale ranging from “1 – not knowing will not harm a patient” to “5 – not knowing will cause possible death to the patient”. A highly relevant item was one which assessed a foundational knowledge or a core skill. A less relevant item assessed on knowledge or skill which was good to know or acquire but not required for progression to the next level of education. The difficulty of an item was rated on a 5-point scale ranging from “1 – very easy” to “5 – very difficult”. The difficulty of the item was dependent on the ease of understanding the item construction or the difficulty of the disease condition assessed. For instance, the inclusion of multiple comorbidities in the item stem, as opposed to one comorbidity, required the student to synthesise information before responding. The difficulty of the item

was also associated with the level of learning that was assessed. Hence, an item which was assessed on application would be more challenging to the student compared to an item assessing recall. The frequency of an item was rated on a 4-point scale from “1 – very rarely seen in practice of a basic doctor” to “4 – seen very often in practice of a basic doctor”. For example, in the local context, influenza is a clinical condition commonly seen in clinical practice while tetanus is a rarer clinical condition.

Judge’s ratings of each criterion were converted into a probability estimate that a borderline candidate would get the item correct ranging from 0 to 100 percent for each item. An item with a low relevance and frequency but a high difficulty would be assigned a probability estimate between 0 to 30 percent suggesting that a borderline candidate was less likely to get the item correct. An item with a high relevance and frequency but a low difficulty would be assigned a probability estimate between 70 to 100 percent suggesting that there was a high probability a borderline candidate would get this item correct. Judges were given the freedom to assign an estimate from the range provided in the guide or to assign a probability estimate based on their own judgement or expertise.

III. IMPLEMENTATION

To date, the guide was shared with judges during the Angoff standard setting sessions for the medical undergraduate assessments. The guide was given at the start of the session when calibrating judges to a similar mental model on what a borderline candidate was. Judges were free to use the guide in the decision-making process when providing a probability estimate for each item. During the calibration phase and discussion phase of the Angoff standard setting session, we observed that judges provided justifications for their probability estimates by referring to the three criteria. This was more prevalent among judges who were new to the Angoff method. We believed that the well-defined and objective criteria provided in the guide served as a useful framework for judges to develop a mental model on what a borderline candidate was.

IV. LIMITATIONS AND FUTURE DIRECTIONS

Several limitations have been identified. While we have attempted to implement the guide, judge’s ratings remained influenced by their own criteria set by their personal experiences and beliefs which were often deeprooted and independent of the three identified criteria. This is especially so for judges who had prior experience in standard setting with Angoff method and had formed their own set of criteria. We see greater value

in the use of the guide for training judges who were participating in Angoff standard setting for the first time.

The guide was developed within a specific medical school in Southeast Asia with its own unique curriculum and learning objectives. Its applicability and effectiveness may be limited in different educational contexts with varying curricula and assessment methods. These limitations highlighted the need for ongoing evaluation and adaptation of the guide and standard-setting methods to ensure they meet the needs of diverse educational settings and provide reliable assessment outcomes. The team is working on validating the use of the guide in our own local context. This would be conducted by quantifying the level of agreement between judges’ ratings, correlating with other standard setting methods and soliciting feedback from judges on the utility of the guide.

V. CONCLUSION

As more medical schools begin to adopt criterion referenced standard setting methods to set a defensible pass mark for assessments and given the complex process judges face when rating items, there is value in the provision of a guide to judges with defined criteria to facilitate the process of rating items.

By focusing on criteria such as relevance, frequency, and difficulty, the guide aimed to provide a structured framework for judges to make more consistent and objective probability estimates of a borderline candidate’s performance. Preliminary observations suggested that the guide has been useful in standardising judges’ evaluations and aligning them with the intended competence levels of a borderline candidate. However, variability in judges’ personal criteria and context-specific development posed potential issues. Pilot testing, inter-rater reliability studies, and expert reviews were essential in evaluating the guide’s impact on the pass marks. Ultimately, a well-validated guide has the potential to improve the fairness and reliability of assessments in medical and health professions education, ensuring that graduating students are competently prepared to provide safe and effective patient care.

Notes on Contributors

Han Ting Jillian Yeo contributed to writing and editing the manuscript.

Dujeepa Samarasekera contributed to the concept and development of the manuscript.

Ethical Approval

No ethical approval was required for this study as no data were collected.

Funding

No funding sources are associated with this paper.

Declaration of Interest

There are no conflicts of interests related to the content presented in the paper.

References

Norcini J. J. (2003). Setting standards on educational tests. *Medical Education*, 37(5), 464–469. <https://doi.org/10.1046/j.1365-2923.2003.01495.x>

Tavakol, M., & Dennick, R. (2017). The foundations of measurement and assessment in medical education. *Medical Teacher*, 39(10), 1010–1015. <https://doi.org/10.1080/0142159X.2017.1359521>

Taylor, C. A., Gurnell, M., Melville, C. R., Kluth, D. C., Johnson, N., & Wass, V. (2017). Variation in passing standards for graduation-level knowledge items at UK medical schools. *Medical Education*, 51(6), 612–620. <https://doi.org/10.1111/medu.13240>

Yeates, P., Cope, N., Luksaite, E., Hassell, A., & Dikomititis, L. (2019). Exploring differences in individual and group judgements in standard setting. *Medical Education*, 53(9), 941–952. <https://doi.org/10.1111/medu.13915>

*Han Ting Jillian Yeo
10 Medical Drive,
Singapore 117597
Email: jillyeo@nus.edu.sg