OPEN ACCESS

Saudi Journal of Biomedical Research

Abbreviated Key Title: Saudi J Biomed Res
ISSN 2518-3214 (Print) |ISSN 2518-3222 (Online)
Scholars Middle East Publishers, Dubai, United Arab Emirates
Journal homepage: https://saudijournals.com

Original Research Article

Familiarity and use of Standard Setting Methods for Multiple Choice Questions among Clinical Dental Educators in Nigeria

Yarhere Kesiena Seun[1]*, Umanah Ayamma Udo[2]

[1]Doctor, Department of Oral and Maxillofacial Surgery, Faculty of Dentistry, College of Health Sciences, University of Port Harcourt, Rivers State, Nigeria
[2]Doctor, Department of Conservative Dentistry, Faculty of Dentistry, University of Port Harcourt, Rivers State, Nigeria

DOI: https://doi.org/10.36348/sjbr.2026.v11i03.003 | **Received:** 16.01.2026 | **Accepted:** 09.03.2026 | **Published:** 16.03.2026

***Corresponding author:** Yarhere Kesiena Seun
Doctor, Department of Oral and Maxillofacial Surgery, Faculty of Dentistry, College of Health Sciences, University of Port Harcourt, Rivers State, Nigeria

Abstract

Validated standard-setting methods for multiple choice questions in dental education will ensure fairness and defensibility. However, many tertiary institutions in Nigeria still depend on arbitrary cut-off scores. Written assessments in dental and medical education increasingly rely on multiple-choice questions (MCQs) that target higher levels of Bloom's taxonomy rather than simple recall. This study explored the familiarity, training and current practices of clinical dental teachers regarding established standard-setting approaches and examined their Angoff judgments for two clinical vignette MCQs designed to assess higher-order cognition. A descriptive cross-sectional electronic survey was completed by 65 clinical teachers across major dental specialties. Sample reflected established teachers rather than novices. Many respondents reported low familiarity with standard setting method particularly for the Ebel, borderline, contrasting groups and blueprinting methods.44.6% reported attending workshops, courses or seminars on standard setting or MCQ construction, whereas 55.4% reported only self-study or no training at all. The pattern across methods suggests that formal training is particularly impactful for familiarity with the Angoff method. Challenges identified included time constraint, insufficient training opportunities, limited access to materials and software, and absence of formal standard-setting panels. Participants expressed strong interest in workshops, guidelines, mentorship, and analytical tools. The findings highlight critical gaps in the application of standard-setting methods in Nigerian dental education and underscore the need for institutionalised faculty development, clearer definitions of borderline competence, and routine psychometric analysis to enhance the validity and defensibility of MCQ-based assessments.

Keywords: Standard Setting, Dental Education, Multiple Choice Questions (MCQs), Angoff Method, Ebel method, blueprinting, Borderline group method, Bloom's taxonomy.

INTRODUCTION

The foundation of quality assurance in undergraduate medical and dental education lies in robust, defensible assessment systems (McLachlan *et al*., 2021) (Hassan, 2024) (Ho *et al.*, 2025a). Standard setting defines the minimum threshold a student must meet to be deemed competent. It ensures fairness, credibility, and alignment with educational outcomes. In Nigerian educational systems (Ekekezie *et al*., 2024), standard setting is often misunderstood, with high-stakes examinations relying heavily on arbitrary cut-off scores, especially in institutions lacking robust faculty development programmes.

High-stakes written assessments in dental and medical education increasingly rely on multiple-choice questions (MCQs) that target higher levels of Bloom's taxonomy such as application, analysis and synthesis rather than simple recall (Liu *et al*., 2024) (Newton, 2024) (Gottlieb *et al*., 2023). To ensure defensible pass/fail decisions, many assessment bodies recommend criterion-referenced standard-setting methods, including the Angoff, Ebel, borderline group and contrasting groups approaches, often supported by blueprinting and psychometric analysis.

Despite these recommendations, many faculties—particularly in low- and middle-income contexts—continue to rely on norm-referenced or arbitrary cut-off scores, with variable staff development in formal standard-setting processes. In such settings, clinical teachers in dentistry may construct MCQs and

judge their difficulty without structured training, raising questions about the consistency of cut-off scores and alignment with intended learning outcomes.

The present study was undertaken in dental faculties where clinical teachers from multiple sub-specialties routinely construct MCQs for undergraduate examinations. The study aimed to (1) describe teachers' familiarity with, and training in, standard-setting methods for MCQs, (2) explore their current MCQ construction and review practices, (3) document perceived challenges and support needs, and (4) obtain Angoff standard-setting judgements for two complex clinical MCQs designed to assess higher-order reasoning.

METHODS

Study design and setting

A descriptive cross-sectional survey was conducted among clinical teachers involved in undergraduate assessment at dental-led medical faculties. Participants represented a range of specialties including Oral and Maxillofacial Surgery, Oral pathology, Oral Medicine, Paediatric dentistry, Restorative dentistry, Orthodontics and Preventive dentistry.

Data collection procedure

The survey was administered electronically using a structured form. The data file includes timestamps for completion, enabling verification that responses were collected over several days between May 1st and July 31st, 2025. Participation was voluntary, and submission implied consent; responses were anonymised at the point of analysis.

Participants and sampling

Eligible participants were lecturers and consultant-level clinicians who (i) were actively involved in teaching and assessment of dental students and (ii) had experience in constructing, reviewing and assessing MCQs for institutional examinations. The survey captured age (categorised as 30–39, 40–49, 50–59, ≥60 years), gender, years in practice (less than 5, 5–10, 11–15, more than 15 years), and specialty. All eligible staff were invited by sending the link to a google form to all dental faculty platforms in Nigeria, and all who completed the online questionnaire within the study period were included in the analysis.

Instrument development

The questionnaire was developed to capture four main domains:

1. **Familiarity with standard-setting methods:** Angoff, Ebel, borderline group, contrasting group and blueprinting, each rated on a Likert scale from "Not familiar =1" through "Familiar = 2" to "Very familiar = 3". Using this formula, mean familiarity was computed:

$$\text{Mean Familiarity} = \frac{(1\times n_{\text{Not familiar}})+(2\times n_{\text{Familiar}})+(3\times n_{\text{Very familiar}})}{N};$$

Where n represents the number of respondents in the category and N is the total number of respondents. Similar methods were used for all linear scale questions in the study.

2. **Training and confidence:** previous exposure to workshops, courses, seminars or self-study on standard setting, and self-rated confidence in applying standard-setting procedures to determine differences in familiarity and practice in standard setting for MCQs.
3. **MCQ construction and review practices:** frequency of designing MCQs targeting higher levels of Bloom's taxonomy, use of recognised good practices (e.g. writing clear stems, avoiding "all of the above/none of the above", ensuring plausible distractors, aligning questions with learning objectives, peer review), use of tools or guidelines, and frequency of difficulty/discrimination analysis.
4. **Perceptions of challenges and support needs:** open-ended items on challenges, and multiple-response options for desired support such as workshops, access to guidelines/manuals, collaboration with experienced colleagues and software tools for MCQ analysis.

To explore practical Angoff standard setting, two clinical vignette MCQs were embedded in the survey. Question 16 described a 12-year-old girl with reduced urine output, generalized oedema and electrolyte abnormalities, and asked respondents to identify the most likely diagnosis from five options. Question 18 described a 16-year-old boy with a painful swelling in the lower jaw, ulcerated oral lesion, systemic features and haematological abnormalities, and asked which therapeutic intervention was most urgently needed. For each, participants indicated the Bloom's level they believed the question assessed (e.g. comprehension, application, analysis, synthesis) and provided an Angoff estimate of the percentage of borderline or "average" students expected to answer correctly.

The instrument items, including the wording of challenges and support options, were refined through internal review by a senior educator, Prof Yarhere, I.E., familiar with assessment but formal pilot data are not included in the dataset. The pilot study was conducted during a faculty of clinical sciences education session in the University of Port Harcourt that included senior faculty members who are experts in setting clinical examinations. The 2 clinical scenarios were modified to reduce ambiguity in the contents and context so that only one option was correct in either of them.

Variables and data handling

Key variables analysed included:

- Demographics: age category, gender, years of practice, and specialty were used. Familiarity with each standard-setting method, categorised as "Not

familiar", "Familiar" or "Very familiar", with an additional "Blueprinting" familiarity item was also created.

- Training exposure: dichotomised as any formal training (workshops, courses, seminars) versus none or self-study only.
- MCQ practices (yes/no): whether respondents used each of the following: clear stems, avoidance of "all of the above/none of the above", plausible distractors, alignment with learning objectives, peer review.
- Use of tools/guidelines (yes/no): any reported use of standard-setting manuals or software tools for MCQ analysis.
- Frequency of psychometric review: self-reported frequency of assessing difficulty and discrimination indices following examinations.
- Presence of a departmental/faculty standard-setting panel (yes/no).
- Perceived challenges were coded from free-text entries and categorised afterwards, thematically.
- Desired support: multi-response selections of workshops, access to guidelines/manuals, collaboration with experienced colleagues and software tools.
- Angoff outcomes used Bloom's level attributed to Question 16 and Question 18, and the estimated percentage of average students expected to answer each correctly.

Data analysis

Descriptive statistics were planned using statistical software. Categorical data (e.g. familiarity categories, training exposure, MCQ practices) were summarised as frequencies and percentages. For the Angoff ratings, mean and standard deviation of the estimated percentage correct were calculated for each vignette, with stratification by specialty group for subspecialties, years of practice and prior formal training in standard setting, where cell sizes were adequate.

Comparisons of proportions were examined using chi-square or Fisher's exact tests, and differences in mean Angoff scores between groups were examined using t-tests or ANOVA, depending on data distribution. Open-ended responses on challenges were grouped thematically into workload/time, limited training and materials, difficulty generating plausible distractors, alignment with objectives and coverage of curriculum, and difficulty assessing item statistics.

Ethical clearance was sought from the institutional review committee prior to data collection.

RESULTS

Participants' characteristics

Sixty-five faculty members across Nigeria responded, giving a response rate of 32.5%, from an expected 200. There were 41 males (63.1%) and 24 females, (36.9%) and subspecialties included Paediatric dentistry, 8 (12.3%), Community dentistry 6 (9.2%), Family dentistry 10 (15.4%), Oral and Maxillofacial surgery 18 (27.7%), Oral pathology 5 (7.7%), Orthodontics 9 (13.85), and Restorative dentistry 9 (13.8%). Thirty-three (50.8%) were in age group 50 – 59 years, 20 (30.8%) were in group 40 – 49 years and the rest 12 (18.5) were older than 60 years.

A substantial proportion of respondents 42, (64.6%) reported more than 10 years in practice and regular involvement in undergraduate teaching and assessment, suggesting that the sample reflected established examiners rather than novices.

Familiarity and training in standard-setting methods

Many respondents reported being "Not familiar" with standard setting methods; Angoff 32 (49.2%), Ebel 44 (67.7%), Borderline 37 (56.9%), Contrasting 38 (58.5%) and Blueprinting 44 (67.7%) Figure 1.

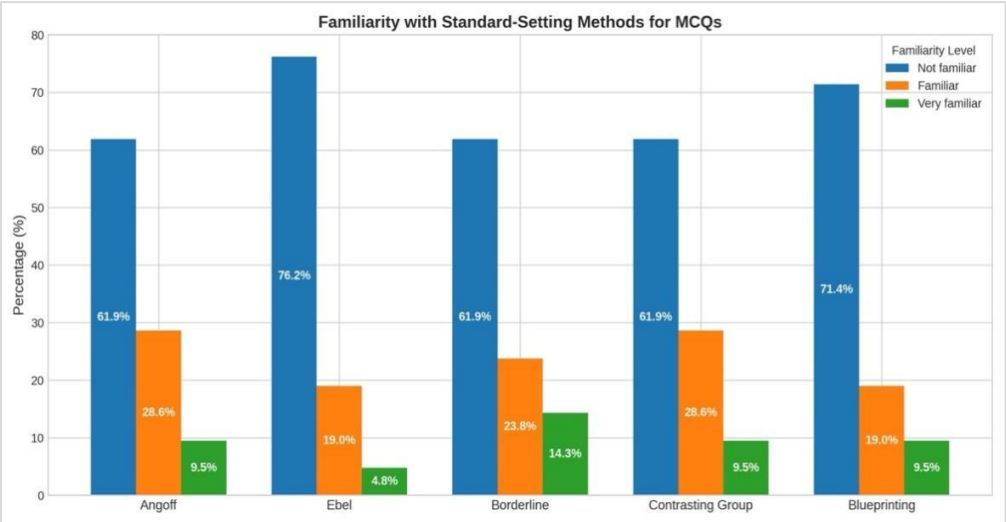

Figure 1: Distribution of respondents' self-reported familiarity with the standard setting methods

Exposure to formal training also varied. Some respondents, 29 (44.6%) reported attending workshops, courses or seminars on standard setting or MCQ construction, whereas others 36 (55.4%) reported only self-study or no training at all. There was a statistically significant association between training status and type of training received ($\chi^2 = 65.00$, $p < .001$), indicating that formal training on standard-setting methods was exclusively delivered through structured institutional programmes.

Familiarity with standard-setting methods was compared between respondents who had received prior training and those who had not, and only Angoff method, had significant association between those with formal training and familiarity of testing method: $\chi^2 = 6.11$, $p = 0.047$, (table I). Trained respondents were more likely to report being "Familiar" or "Very familiar" with the Angoff method, whereas untrained respondents were predominantly "Not familiar." Overall, the pattern across methods suggests that formal training is particularly impactful for familiarity with the Angoff method, with more modest or non-significant effects for the other standard-setting approaches.

Table I: Association between training and familiarity with different standard setting methods

Method	Familiarity Level	FT:Yes (n=29)	FT:No (n=35)	Fisher's	df	p-value
Angoff	Not familiar	11 (39%)	27 (77%)	6.11	2	0.047
	Familiar	14 (48%)	6 (17%)			
	Very familiar	4 (13%)	2 (6%)			
Ebel	Not familiar	17 (59%)	31 (89%)	3.74	2	0.154
	Familiar	10 (34%)	4 (11%)			
	Very familiar	2 (7%)	0 (0%)			
Borderline	Not familiar	14 (48%)	26 (74%)	4.60	2	0.100
	Familiar	11 (38%)	6 (17%)			
	Very familiar	4 (14%)	3 (9%)			
Contrasting Group	Not familiar	14 (48%)	27 (77%)	5.80	2	0.055
	Familiar	11 (38%)	6 (17%)			
	Very familiar	4 (14%)	2 (6%)			
Blueprinting	Not familiar	17 (59%)	31 (89%)	1.69	2	0.429
	Familiar	10 (34%)	4 (11%)			
	Very familiar	2 (7%)	0 (0%)			

FT: Formal training

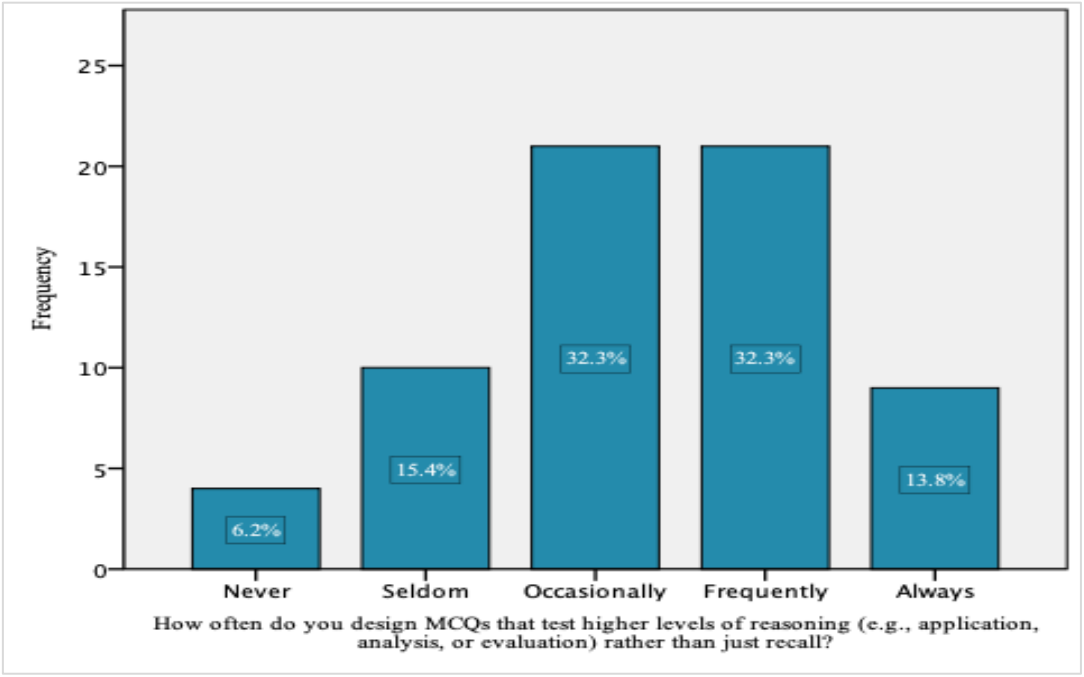

Figure 2a: Bar chart showing frequency with which respondents design higher order MCQs,

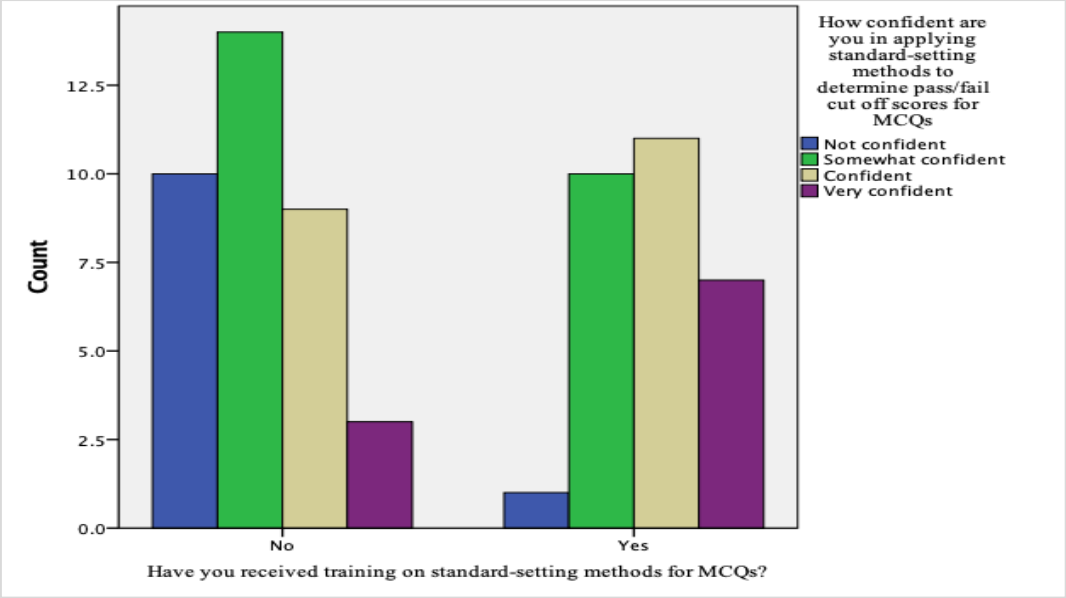

Figure 2b: clustered bar chart comparing training and confidence in designing higher order MCQs.

Thirty (46.1%) respondents design MCQs that test higher levels of reasoning in their practice i.e frequently and always. Respondents who received formal training in standard setting for MCQs were significantly more confident in their ability to set questions than those who did not receive formal training; $\chi^2 = 9.183$, df = 3, p = 0.027 (Figure 2a).

MCQ construction and review practices

Most respondents indicate that they routinely adopt recognised good practices when constructing MCQs, including writing clear and concise stems, aligning questions with learning objectives and ensuring plausible distractors. Many explicitly report avoiding "all of the above/none of the above" options and using peer review to improve question quality, although a subset do not engage in regular peer review.

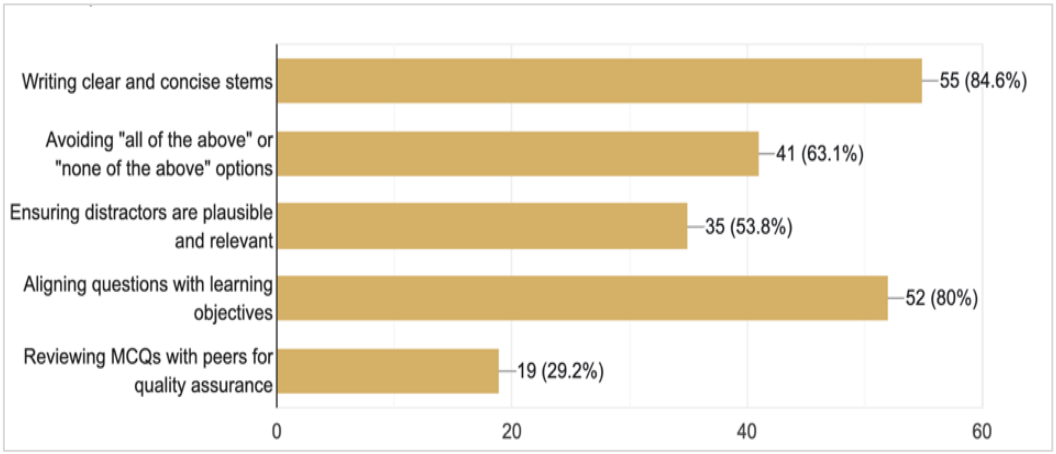

Figure 3: Bar chart showing MCQ construction practices by respondents

Use of specific tools or guidelines is inconsistent. Some respondents mention using standard-setting guidelines or software tools for MCQ analysis, while others report relying solely on personal judgment and experience. A considerable number state that they rarely or never compute difficulty and discrimination indices after examinations, highlighting limited routine psychometric evaluation.

Presence of panels, challenges and support needs

Responses to items on departmental or faculty standard-setting panels indicate that formal panels are not universally present; some departments appear to standard-set informally or individually. Participants describe a variety of challenges, which cluster around:

- Time constraints and the labour-intensive nature of constructing high-quality MCQs and engaging in formal standard setting.
- Lack of formal training, limited access to materials and guidelines and, in some cases, lack of software to conduct item analysis.
- Difficulties in generating plausible distractors, avoiding ambiguity and aligning items with a

broad curriculum and specific learning objectives, particularly when trying to test higher-order reasoning rather than recall.

- Practical constraints such as examiner availability, shortage of manpower and, occasionally, resistance from colleagues to changing established practices.

When asked about support that would improve their skills, respondents frequently select workshops or training sessions, access to standard-setting guidelines or manuals, opportunities for collaboration with experienced colleagues and software tools for MCQ analysis. These responses suggest strong interest in structured capacity-building and institutional support.

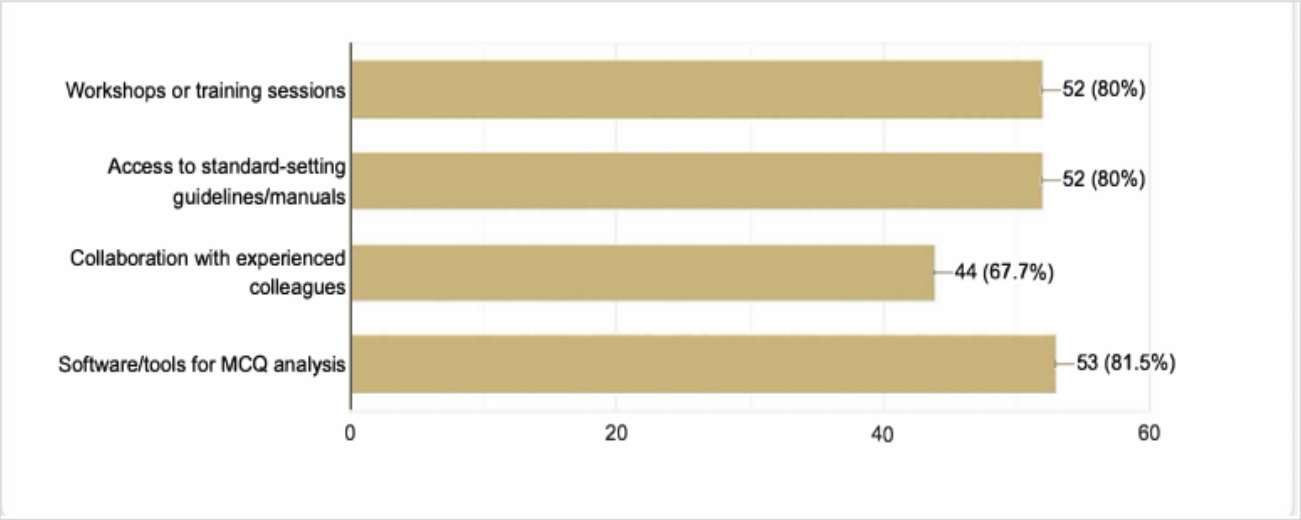

Figure 4: Chart showing support that would improve skills

Angoff judgements for clinical vignettes

For the clinical vignette in Question 16, 76.9% of respondents classified this as application or higher, out of which, 44.6% agreed it assessed application correctly. The Angoff estimates showed that there are 2 different interpretations of difficulty, i.e., Group A, (56% + 48% ≈ 55% of respondents. These respondents think only about half of the average students will get the answer correctly, suggesting that examiners view this as a moderately difficult question for borderline candidates. On the other hand, Group B, (75% + 90% + 96% ≈ 45% of respondents think most students will get the question correct, perceiving it as moderately easy.

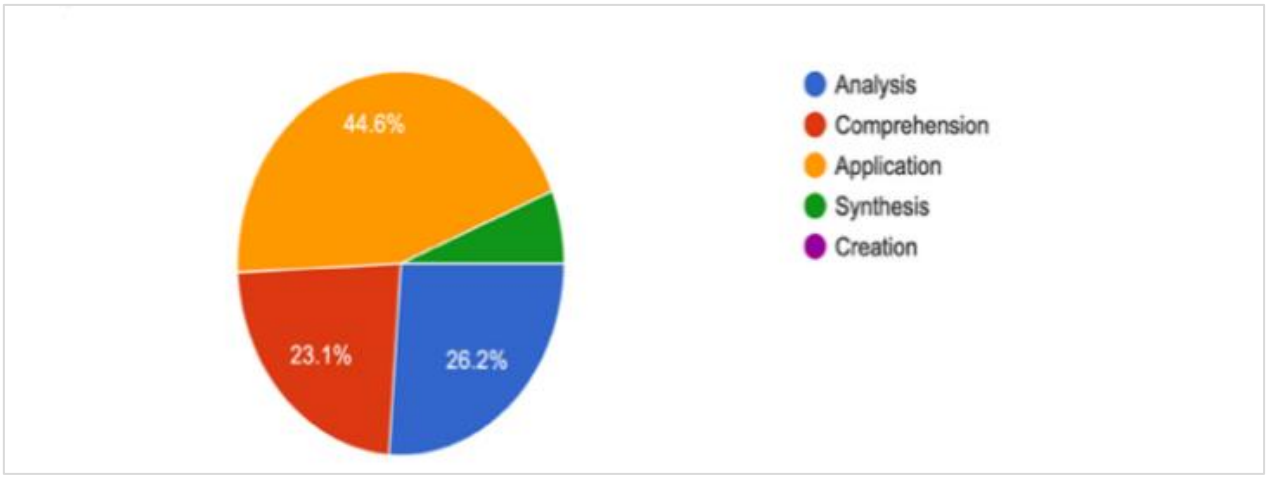

Figure 5a: Pie chart showing how respondents classified the cognitive level of the MCQ

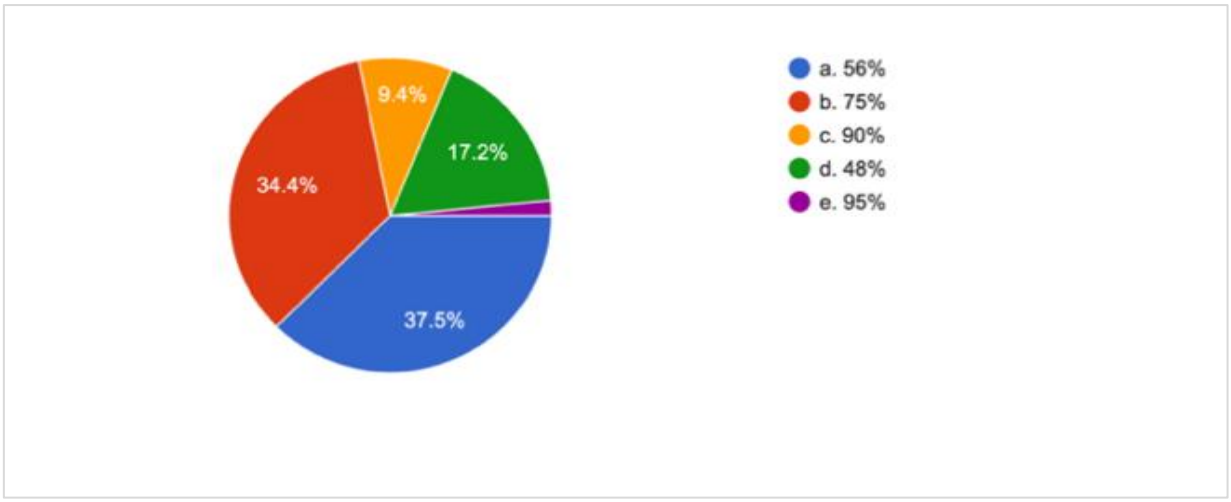

Figure 5b: Pie chart showing respondents thoughts on what percentage of average students will answer the question correctly

For Question 18, a greater proportion 84.6% classified the vignette as application or higher, with 27% correctly assigning analysis to it. The Angoff estimates showed that 2 dominant estimates i.e., (56% + 48% cluster around 50 – 60% of respondents. These respondents think only about half of the average students will get the answer correctly, suggesting that examiners view this as a moderately difficult question for borderline candidates.

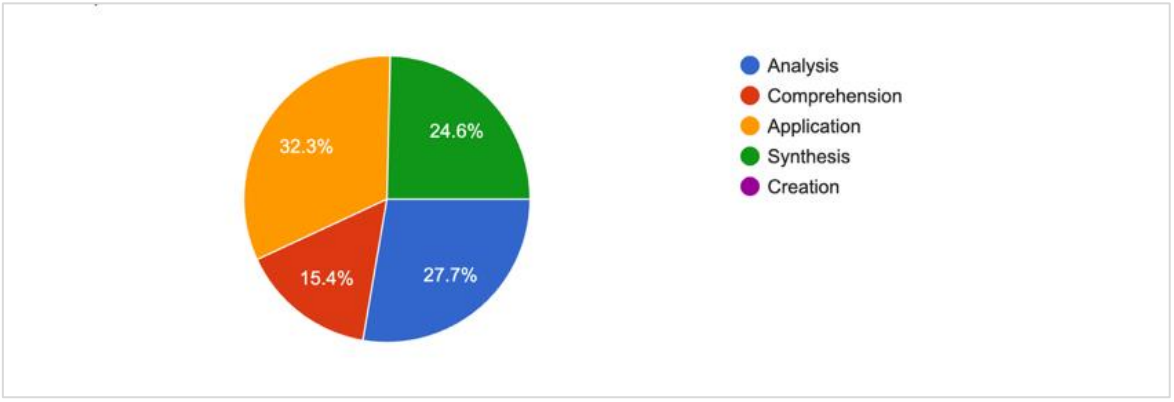

Figure 6a: Pie chart showing how respondents classified the cognitive level of the MCQ

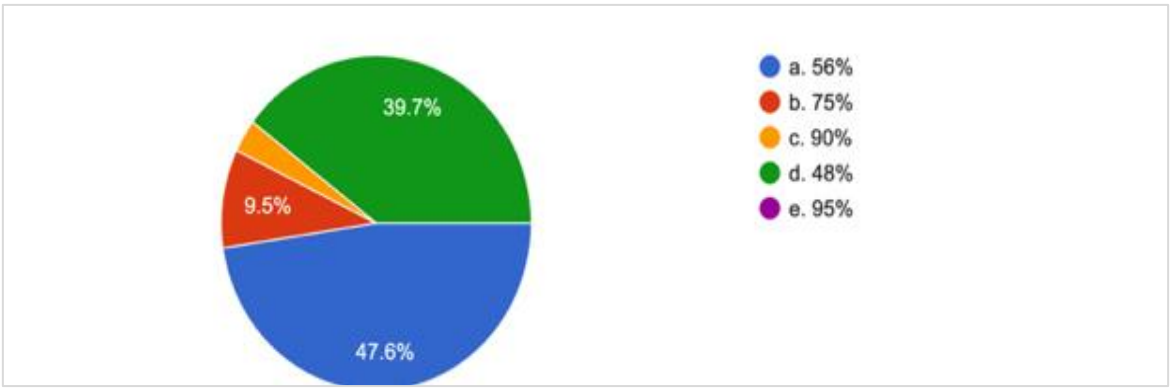

Figure 6b: Pie chart showing respondents thoughts on what percentage of average students will answer the question correctly

Across both questions, respondents' judgements were based on the length of the vignette (41 – 45%), that time to make decisions in examinations was short 11.5%.

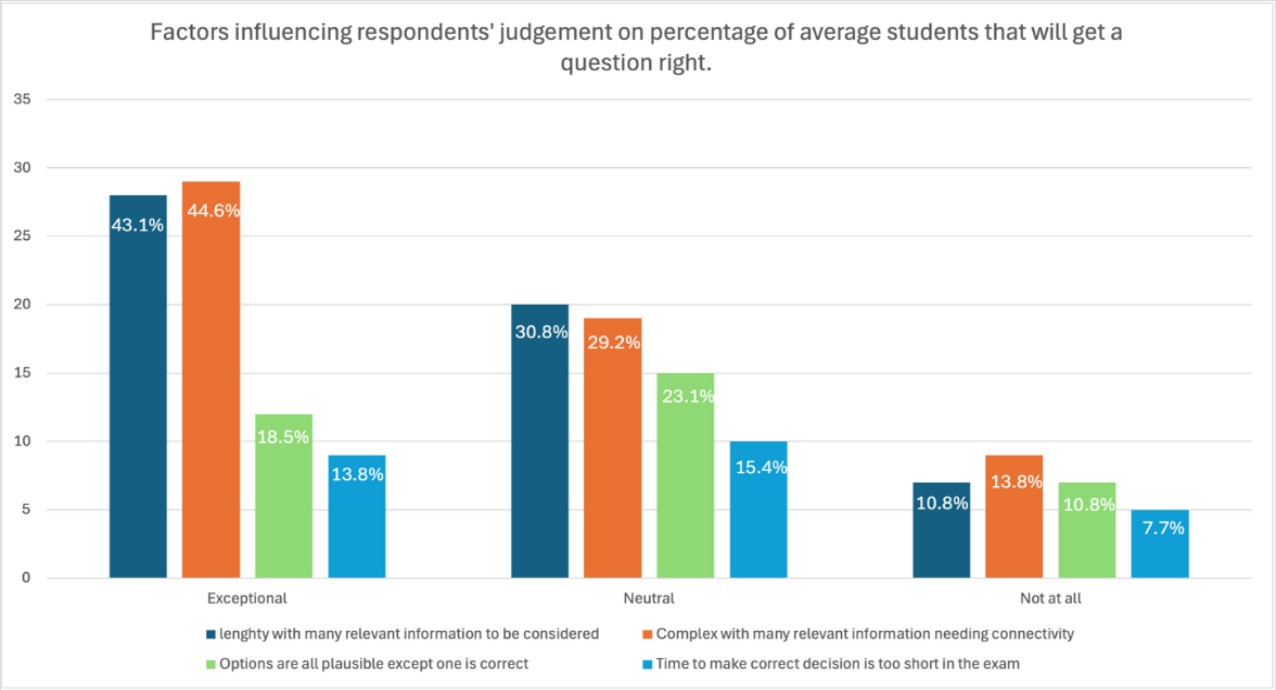

Figure 7: Percentage distribution of factors influencing respondents' judgements of MCQ difficulty

DISCUSSION

This study has shown that most senior and experienced Dental faculty members have low familiarity with established standard setting methods for MCQs, especially Ebel, and blueprinting. This stands to reason as more than half the respondents also agree that they have not received formal training in standard setting. A comparison of familiarity levels between respondents who have received formal training and those who have not showed significant differences. Nearly three quarters of the responses from those yet to have formal training fell within the not familiar category. A 2017 (Ho *et al*., 2025a) paper pointed out that many dental faculty enter academia without formal training in education giving this gap in standard setting. Another paper argues that without proper training in these methodologies, the resulting cut-scores will lack the defensibility and validity that is required for minimum competence for patient safety.(Hassan, 2024) This is burdensome as there are possibilities that these faculty members will continue in their role of teaching and assessment without dedicated pedagogical training as clinical practice will take centre stages in their roles. Many respondents were familiar with Angoff method of standard setting and this was seen more in persons with formal training, which is consistent with practices worldwide as Angoff is most popular method with some experts modifying it to suit their unique situations. (Ho *et al*., 2025a) (Dimanche *et al*., 2025)

The finding that most participants routinely adopt good MCQ writing practices (clear stems, plausible distractors, alignment with learning objectives, avoidance of "all of the above/none of the above") is encouraging and aligns with international guidance on item writing. However, respondents attributed difficulties with lengthy, information dense stems and complex vignette are consistent with cognitive load theory, which informed the choices of level of difficulty in the vignettes. (Chen *et al*., 2023)Though plausible distractions are known to increase item difficulty and discrimination, (Lee *et al*., 2025) respondents tend to weigh stem complexities heavier than distractors plausibility in their Angoff judgements. The time pressure perceived as reasons for vignette difficulty aligns with item writing guidance which emphasizes stem clarity, cognitive demands and distractor quality as primary levers for controlling difficulties, (Excellence, 2019) Perceived challenges—especially time constraints, lack of training and materials, difficulties generating plausible distractors and limited examiner availability—are consistent with barriers documented in many resource-constrained educational settings. (Zhang, 2021) (Abd-Rahman *et al*., 2021; Brown *et al*., 2021)

The strong expressed desire for workshops, access to guidelines, collaboration with experienced colleagues and software support indicates both recognition of these gaps and readiness to engage in capacity-building initiatives. *(Ho et al*., 2025b) Training was also associated with greater confidence in constructing higher-order MCQs and in applying standard-setting methods. The association is consistent with quasi-experimental studies showing that structured faculty-development programmes improve MCQ writing quality, increase higher-order item production, and reduce item-writing flaws. (Abdulghani *et al*., 2015) Faculty development interventions have reduced non-functioning distractors and item-writing flaws,

resulting in better psychometric indices and student performance. (Liu *et al*., 2024; Newton, 2024)

The Angoff ratings for the two clinical vignettes provide insight into how clinicians operationalise standard setting in practice. Respondents generally recognised the vignettes as application or analysis item. However, for the 2 questions, though fewer than 50% judged them correctly, they were appropriately given their complexity. Nonetheless, the observed spread of Angoff estimates across respondents suggests variability in how borderline or "average" performance is conceptualised. The bimodal Angoff distribution for Question 16 plausibly reflects differing mental models of the "borderline" student and variability in examiners' experience with similar items. Such variability underscores the need for structured panel discussion, calibration exercises and explicit definition of the borderline student when applying Angoff and related methods in high-stakes examinations. Disagreement in Angoff estimates may indicate inadequate training or insufficient group discussion rather than genuine differences in expectations about student performance. (Abdulghani *et al*., 2015) Reviews of Angoff methods emphasise that panel training, clear definition of the borderline candidate, and iterative "reality checks" are essential to reduce variability and improve reliability of cut scores. As there was no focus group discussion in the survey, alignment of estimates by respondents will be difficult and this should be taken into consideration in future research. Not many understand who the average student is and many faculty expect standards higher than the students' levels. Research on faculty perception of item difficulty shows that without empirical feedback, judges' estimates can be biased by superficial item characteristics such as vignette length or perceived complexities. (McLachlan *et al*., 2021)

Across both vignettes, many respondents reported basing their difficulty judgements on factors such as vignette length (41–45%) and perceived time pressure during exams (about 11.5%). These heuristic cues may overshadow a more analytic consideration of content complexity, prerequisite knowledge, and common student misconceptions. Vignette length and time pressure do genuinely contribute to intrinsic and extraneous cognitive load, so they are not entirely irrelevant to difficulty estimation. Strengthening formal training in standard-setting methods, instituting regular departmental or faculty-wide standard-setting panels and embedding routine psychometric analysis could significantly enhance the validity, fairness and transparency of MCQ-based assessments for dental students. (Hassan, 2024) (Ho *et al*., 2025a) (Ho *et al*., 2025b) (Ward *et al*., 2018)

Implications for practice

The findings support several practical recommendations:

- Institutionalize regular workshops and refresher courses on Angoff, Ebel and related methods, prioritizing staff who currently report no or minimal training.
- Develop and disseminate concise standard-setting and item-writing guidelines, tailored to local curricula, with exemplar MCQs at different Bloom levels.
- Establish multidisciplinary standard-setting panels that include dental and non-dental specialists, with structured calibration before applying Angoff judgements.
- Implement user-friendly software for difficulty and discrimination analysis and incorporate item statistics into post-exam review routines.

Strengths and limitations

A strength of this study is the inclusion of a wide range of clinical and senior examiners across dentistry, providing a broad snapshot of standard-setting knowledge and practices within one faculty. The integration of two clinical vignette MCQs with Angoff ratings adds a practical dimension, linking self-reported familiarity to real standard-setting judgements.

Limitations include reliance on self-reported familiarity and training, which may be subject to social desirability and recall bias, and the absence of detailed psychometric data for the vignettes themselves. This may also be influenced by the desire to portray competence, especially since participants are senior clinicians, so we recommend triangulation with objective data in future research.

Acknowledgement

We wish to acknowledge Prof. Iroro E. Yarhere for helping with the face and content validity of the survey questionnaire and guidance in the conceptual and theoretical framework for this study. Our sincere appreciation goes to Prof. O.A Akadiri for his invaluable contribution to this study.

REFERENCES

1. McLachlan JC, Robertson KA, Weller B, Sawdon M. An inexpensive retrospective standard setting method based on item facilities. *BMC Med Educ.* 2021;21(1):7. doi:10.1186/s12909-020-02418-5.
2. Hassan S. Standard setting in medical education: Standards, Methods, and Psychometrics. Global Medical Education in Normal and Challenging Times. Springer; 2024. p. 137-150.
3. Ho TK, Abu Kassim NL, O'Malley L, Roudsari RV. Standard setting for dental knowledge tests: reproducibility of the modified Angoff and Ebel method across judges. BMC Med Educ. 2025;25(1):1426. doi:10.1186/s12909-025-07822-3.
4. Ekekezie OO, Charles-Eromosele TO, Olatona FA, Aguwa EN. Standard-setting methods for assessment in a post-graduate medical college. Niger Postgrad Med J. 2024;31(3):263-268. doi: 10.4103/npmj.npmj_72_24.

5. Liu Q, Wald N, Daskon C, Harland T. Multiple-choice questions (MCQs) for higher-order cognition: perspectives of university teachers. Innov Educ Teach Int. 2024;61(4):802-814. doi:10.1080/14703297.2023.2222715.
6. Newton PM. Guidelines for creating online MCQ-based exams to evaluate higher order learning and reduce academic misconduct. In: Eaton SE, editor. Handbook of academic integrity. Singapore: Springer; 2023. doi:10.1007/978-981-287-079-7_93-1.
7. Gottlieb M, Bailitz J, Fix M, Shappell E, Wagner MJ. Educator's blueprint: a how-to guide for developing high-quality multiple-choice questions. AEM Educ Train. 2023;7(1): e10836. doi:10.1002/aet2.10836.
8. Dimanche K, Klatt EC Jr, Angle SM. Predictive validity evidence of Yes-No Angoff standard setting in a pre-clinical medical school curriculum. BMC Med Educ. 2025;25(1):384. doi:10.1186/s12909-025-06948-8.
9. Chen O, Paas F, Sweller J. A cognitive load theory approach to defining and measuring task complexity through element interactivity. *Educ Psychol Rev.* 2023; 35:63. doi:10.1007/s10648-023-09782-w.
10. Lee Y, Kim S, Jo Y. Generating plausible distractors for multiple-choice questions via student choice prediction. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; 2025; Vienna, Austria. Association for Computational Linguistics. p. 23669–23692.
11. Excellence SIFT. Guide to item analysis. Pennsylvania State University; 2019.
12. Zhang X. Stakeholders' test perceptions on test reform. Stud Educ Eval. 2021;70: 101064. doi:10.1016/ j.stueduc. 2021.101064
13. Abd-Rahman ANA, Baharuddin IH, Abu-Hassan MI, Davies SJ. A comparison of different standard-setting methods for professional qualifying dental examination. *J Dent Educ.* 2021;85(7):1210-1216. doi:10.1002/jdd.12600.
14. Brown G, Denny P, San Jose D, Li E. Setting standards with multiple-choice tests: a preliminary intended-user evaluation of SmartStandardSet. *Front Educ.* 2021; 6:735088. doi:10.3389/feduc.2021.735088.
15. Ho TK, O'Malley L, Roudsari RV. Investigating assessment standards and fixed passing marks in dental undergraduate finals: a mixed-methods approach. *BMC Med Educ.* 2025;25(1):481. Published 2025 Apr 3. doi:10.1186/s12909-025-06944-y.
16. Abdulghani HM, Ahmad F, Irshad M, *et al*. Faculty development programs improve the quality of multiple-choice questions items' writing. *Sci Rep.* 2015;5 :9556. doi:10.1038/srep09556.
17. Ward H, Chiavaroli N, Fraser J, *et al*. Standard setting in Australian medical schools. *BMC Med Educ.* 2018;18(1):80. Published 2018 Apr 23. doi:10.1186/s12909-018-1190-6.