frontiers | Frontiers in Medicine

TYPE Opinion
PUBLISHED 30 November 2023
DOI 10.3389/fmed.2023.1287924

OPEN ACCESS

EDITED BY
Haider Al-Waeli,
Dalhousie University, Canada

REVIEWED BY
Soledad Armijo,
San Sebastián University, Chile

*CORRESPONDENCE
Avita Rath
drrathavita@yahoo.com

RECEIVED 03 September 2023
ACCEPTED 06 November 2023
PUBLISHED 30 November 2023

CITATION
Rath A (2023) Back to basics: reflective take on role of MCQs in undergraduate Malaysian dental professional qualifying exams.
Front. Med. 10:1287924.
doi: 10.3389/fmed.2023.1287924

Back to basics: reflective take on role of MCQs in undergraduate Malaysian dental professional qualifying exams

Avita Rath [1,2]*

[1]Faculty of Dentistry, SEGi University, Petaling Jaya, Selangor, Malaysia, [2]Edinburgh Medical School-Clinical Education, University of Edinburgh, Edinburgh, United Kingdom

KEYWORDS

dental education, assessment, MCQ assessment, high-stake tests, undergraduate

Introduction

The Bachelor's Dental Programme (BDS) in Malaysia is a 5-year full-time undergraduate course, the tenets of which lie in an overarching competency- and outcome-based curriculum (1). It aims to prepare dental students to become independent, reflective practitioners who deliver quality patient care (2). The programme aims at organizing the graduates' attributes around a wide range of competencies that include evidence-based knowledge, critical thinking, problem-solving, procedural skills, ethical values, and professionalism (3, 4). It also emphasizes student-centered learning and provides a design-down framework based on attainable learning objectives that drive the pedagogy/instructions reflected in an authentic assessment (5, 6).

The final summative assessment, or the professional examination, is usually a combination of written and performance-based formats that aim to measure the different facets of competencies in alignment with the course goals as per the Malaysian Dental Council guidelines (3). Written assessments that entail 60% of the final grades have long prevailed in assessments to capture the cognitive domains of Bloom's Taxonomy that may span from knowledge recall to evaluation or capture the "knows and knows-how" of Miller's pyramid of competency (4, 7, 8). Among other formats, multiple choice questions (MCQ) is the most sought-after design for these forms of assessments. Albeit known for their ubiquitous presence due to their testing breadth of knowledge and ease of administration, they are spuriously known to defy recommended guidelines and have garnered a negative reputation for engaging lower cognitive domains or even the test-wiseness ability in lieu of actual knowledge (9, 10). Our existing MCQ paper consists of 60 items of one-correct answer (OCA) with four options and complex two-tier or K-type questions that predominantly assess rote recall. Supposedly, if the final summative assessment provides legitimacy by certifying the measured competencies (11), in that case, the predictive accuracy of an assessment toward measured competency (12) may be questionable, putting the quality of the entire programme at risk and prompting immediate action (13). Moreover, under the new dental act (14), the graduates must appear for a professional qualifying examination (PQE), a licensing exam with single best answers (SBA), and an objective structured clinical examination (OSCE) format commencing in 2025, to register for practice (14). Hence, it seems incumbent to go back to basics, revisit MCQ for its worth as an authentic assessment tool, and take a pragmatic approach contingent on its pros and cons, its acquiescence with other assessment formats, and its fitness for the purpose of qualifying exams for courses like dentistry such as ours.

Purposes of the assessment

Boud famously stipulated that "assessment always does double duty" (15). Based on the stakes involved, those duties/purposes could be broadly divided as formative or assessment for learning, which are low-stakes, ongoing, address the gaps, and assimilate notions construed by the learners by re-clarifying the learning outcomes. Whereas summative or assessment of learning is applied at the end of a module or the course itself. It forms the crux for high-stakes decisions to pass or fail. The data accrued from these assessments further typifies the programme evaluation and holds accountability to stakeholders (16).

In reality, there is always "a continuum of summative to formative…, depending on the primary intended purpose" (17). Therefore, the goals of an assessment tool are contingent on its purpose, which influences its content and strategies (18).

Any assessment tool is informed by a fair share of strengths and weaknesses (19). Hence its utility (U) or usefulness, a conceptual layout posited by van der Vleuten, is a function of the prescribed criteria of reliability (R), validity (V), cost (C), acceptability (A), and educational impact (E), wherein the *weighting* (w) of each component is akin to the purpose of the assessment (20).

$$U_W = R_W \times V_W \times C_W \times A_W \times E_W.$$

Thus, from the vantage point of this enduring framework and other literature is the inductivist way to critically appraise the purpose of the revised tool (21).

An indispensable criterion of high-stakes assessment is the reliability or reproducibility of the scores (20), which is also associated with the validity of its internal structure (17). Often expressed as Cronbach's alpha coefficient (α; ranging from 0 to 1), the values above 0.8 are deemed acceptable for high-stakes exams (22).

Evidence suggests that MCQ formats are renowned for their high reliability (23–25). A common misconception was that its high reliability was due to its objectivity (26). On the contrary, high reliability is borne out of an adequate sampling of questions and as a function of testing time (20).

A possible suggestion of well-designed SBA items over 2–3 h in place of the current 1.75 h 60-item K-type MCQ paper may demonstrate high reliability, as shown in my recent study in the medical education context (27, 28), and must be considered for future assessments.

Conversely, I would like to highlight the hazards of confining reliability to the numeric α alone. It only expresses the degree of replicability of the rank order of candidates or the internal consistency of the scores (29) and doesn't recognize the discriminating power in performances, a primacy for high-stake decisions (17). It is the discrimination index (DI) that describes the discernment capability of an item to differentiate between scorers based on their proficiency in the tested domain (30). Ranging from −1 to +1, which traditionally corresponds to the top and bottom 27% of the cohort, a DI of ≥0.3 for 50–60 items probably would give good reliability (17). The main enemies of DI are the item writing flaws (IWF), such as the implausible or non-functioning distractors (NFDs), *one* of the common rogues in the existing tool (31). NFDs are an option(s) of a question other than the correct answer, which is generally selected by <5% of the examinees (A-value) and illustrated as trace lines (26).

The handed-down format of previous years with 4- or 5-options has never been investigated for IWFs, indicating a heuristic mentality that more options imply increased difficulty with reduced guessing or cueing effect, which may be true provided there were no IWF (32). Moreover, an 80-year meta-analysis clarifies that it is not feasible for more than three plausible distractors and that it would suffice the DI of an MCQ paper (33). Nevertheless, research has shown that a variable number of options based on their educational availability would bolster the content validity of the item, concurrently strengthening its reliability and underpinning my recommendation (34, 35).

Furthermore, DI is also related to an item's difficulty or facility index (*P*-value), expressed in the range of 0 to 1.0, where a higher *P*-value denotes an easier question. Data entail that the SBA tool should have a moderate range of *P*-values (0.25–0.75) to foster a good DI (30). Having said that, some of the items may be defined by learning outcomes that assess the lower levels of cognition and are ostensibly easy for final-year students. Conversely, too many of these items, as seen in the existing tool, are predisposed to higher IWFs and would deter the high performers, threatening the tool's validity (10, 30, 36).

Alongside reliability, validity is another fundamental attribute of the summative exam, which concerns whether the scores measure the competency it purports to (12, 17). A caveat to note is that reliability is a prerequisite to the validity of an assessment; however, it does not ascertain its validity (37).

Modern concepts of validity are overarching and posit a "unitary" framework based on the premise of the fidelity of scores and their inferences (38, 39). I will be highlighting the pertinent concepts with a few mentions of others within the constraints of the space here. Foremost is based on the content of the assessment tool, which should be constructively aligned with the learning outcomes of the topics (29, 40). This is ensured through blueprinting, a method where the test items are mapped against the relevant learning objectives set at appropriate taxonomic levels prior to the commencement of the academic year (18). It apprehends the threat of construct under-representation (CUR)—under-sampling or oversampling of the course content (41). In spite of an entrenched blueprint in our faculty, CUR issues have been noticed, especially in a theme-based MCQ paper when there is an overcompensation by items from feasible topics or when the existing items that are nominated for higher cognitive levels tend to elicit factual recalls (41). Consequently, I would want to paraphrase Coderre's opinions here, which state that audit adherence to the blueprint is required and that creating it alone is insufficient (42). Every item of the new tool should be evaluated for its accurate representation and suitability of the learning outcome for fairer and more reliable scores (43).

Validity is a nebulous term, especially for the critics of the MCQs, who question the authenticity of this close-ended design in eliciting clinical reasoning, which is more nuanced than just selecting an option (44, 45). I acknowledge the connotations of these arguments; nevertheless, one should bear in mind that the inability to record the reasoning processes does not insinuate their absence (46). Moreover, authenticity is present at all levels of

the pyramid (37). Based on this conclusion, as clinical reasoning requires integrative knowledge that entails high-order cognitive skills of application and analysis (18), the new tool with well-designed clinical vignettes could invoke this domain of human endeavor regardless of the response format (47).

That said, the veracity of the stimulus generated could be eluded by errors or "noise" in items leading to the construction of irrelevant variance (CIV) such as grammatical chicanery, complex language, and pseudo-vignettes in a trivial pursuit of elusive "blueprint alignment" draining its fitness in summative exams (17, 48). With no intentions to gainsay written formats that are susceptible to CIV (23, 49), the feasibility of obviating a CIV in an MCQ is higher owing to its compact design (36, 50).

Facets of validity also converge with other utility criteria and would permeate the next section of this discussion. For instance, the *consequential validity* is somewhat analogous to the educational impact (17). A deep-set reality initially underscored by van der Vleuten was that "*Assessment drives learning*" through its content, format, timing, and feedback (20), especially when the summative culture looms large. One must understand that students are *agentic* learners who always prioritize their learning around exams. My take is to be astute and capitalize on these drivers by focusing on the design choices of the new SBA and its strategic placement within the toolkit that would determine its influences within the precincts of the programmatic assessment (21).

Design choices

It seems axiomatic that the educational impact of an assessment is inextricably linked to the assessment literacy of stakeholders, which might be scarce in my setting. Every student at our faculty (SEGi University) owns a handbook with the layout of the assessment structure. However, there is a lack of emphasis on a meta-dialogue early on about the purposes and function (51), as most faculty are at the outset of the curriculum and assessment (52). Although we have had a few cursory workshops, marshaling nuggets of information, the insights are tentative and might have negative implications for the fairness of this tool. In my view, fairness is more of an annotation of the assessment process itself than a design choice, so it is quixotic to address it. To a great extent, it is associated with the stakeholders' acceptability and other utility criteria (24).

Keeping the good name of the new SBA format requires early intervention at the item development stage to avoid CIVs and CURs, as seen in the previous section. As item writing has always been referred to as an art (53), to improvise and excel, it calls for extensive training for most of us who have an intuitive idea of suboptimal design but lack the acumen to identify it. But that would incur a cost—not just economic, but faculty's time—enshrined beliefs further restrained by university policies. Even so, weighing the cost relative to its purpose (20) asserts training to be a worthy investment in the long haul of superior assessments.

Research also asperses the format for inducing adverse *testing effects* where an incorrect answer choice lures the examinees to recall their facts wrong for other exams (45). Albeit not wholly avoidable owing to the selected response design, the scheduling of the paper might mitigate this issue to some extent. Currently, the MCQ is the last paper, which seems suitable for the revised tool, as the deep learning expended around other formats should generate a positive testing effect (54). However, one can never predict the educational impact of assessments without thorough screening and follow-up (20). In fact, to assess the fitness of the new tool, there need to be qualitative pre-assessment and post-assessment checks.

Albeit a *de-facto* review process that occurs at the subject, faculty, and external examiners' stages, it might fortify the quality assurance of the questions at a speciality and interdisciplinary level (55, 56). There is a possibility to make these sessions more defensible and credible so that assessment practices are more legitimate and a good fit for the cohort and the curriculum by mandating standard-setting and item analyses (57).

Standard is a conceptual boundary on a "true" cut-score scale that differentiates between acceptable and non-acceptable performance; in other words, optimal or passing standards can be viewed as an agreed definition of competence that reflects expert judgement as to what constitutes it, backed by several sources of evidence (58). Based on Kane's view of valid inferences (59), relevancy evaluations of an assessment to a well-mapped blueprint are prerequisites to setting standards. It delineates what a competent student needs to know vs. what they must know about a construct, as cited in Schuwirth and van der Vleuten (60). This allows for setting a cut-off or passing score on an observed score scale that should be used to make a defensible, deliberate judgement for that relevant competence. For an SBA tool, a criterion-referenced or absolute test-centered standard setting such as the modified Angoff method is the most appealing as the judgements are made on individual items based on item analyses in the backdrop of minimal competence. It also gives wiggle room for discussion and consensus around the performance data (61, 62).

Nevertheless, practice exercises in the course revealed that it is not possible without psychometric experts, who could employ the correct model for it. Classical test theory (CTT), which consists of α, *P*-value, DI, and *A*-value, discussed earlier, sits well here due to its uni-dimensional construct and simple statistical software (63). Moreover, judgements can be fallible and time-consuming due to a lack of expertise, so it requires the selection of a judges' panel of every age and gender with knowledge of the curriculum. Those who could articulate characteristics of a "minimally competent" based on the cohort's abilities were at the "borderline" of pass and fail (64, 65).

We currently follow an absolute standard of 50% pass-score based on a compensatory method that combines all the formats to produce average marks translated to grades. Evidence reveals that combining scores across the papers to moderate the errors of individual formats is highly reliable for high-stakes decisions (26). Conversely, this method may induce a minimalist study strategy, wherein past students have passed by doing well in specific papers alone. But this point of view imparts a reductionist approach toward competency. Pioneers drew on these issues and espoused a programmatic approach toward assessment (37) that pleads on the holistic narrative of competency, vying that "any single assessment is a weak data point and implies a compromise on the quality criteria" (20). It is always recommended to deploy a deliberate suite of assessments that ameliorates the trade-offs of the utility of various formats, such that collated information is more than the sum of its parts (19). This principle underlies the assumption of

triangulating data from multiple sources and formats throughout the year based on domain specificity, providing robust, meaningful conclusions toward competency rather than relying on a single format (21).

Since the exam is a recursive process, it is also important to perform post-exam item analyses, which would corroborate the credibility and defensibility of the assessment (66). The exercise would yield meaningful feedback for future pre-assessment analyses, identify errors in unfair scores, and, most importantly, justify the need for remediation through the resit exam. Usually, our faculty allows a single resit opportunity after an exit exam within 2 weeks of the final results; however, the ideal number of attempts is debatable (67). Considering the limited faculty resources, especially with a PQE lurking in a few weeks, a single resit looks like the only option for now. Moreover, the advent of PQE seems promising toward desirable but nearly absent catalytic effect or educational feedback (68) from an exit exam (69).

To paraphrase, the possibility of a "fairy-tale" assessment is the wrong question to start with. The burgeoning assessment literature has revealed that there is no ideal tool as they are not goals in themselves, not even my proposed tool. Nonetheless, despite the format's long pedigree, its (over)usage should be monitored in the context of the programme. In summary, the gargantuan responsibility of assessment tools to credibly answer the relentless inquiry of "how much is good enough?" is an outdated pursuit. Thus, it is no longer a question of measurement but an integral issue of the curricular design and the users' expertise in the organizational culture. Moreover, I would contend that the system of continuous longitudinal assessment must be designed with an attempt to operationalize a programmatic assessment format broadly aligned with principles suggested by the proponents of assessment philosophy as discussed above if we wish that assessment to provide authentic information about our learners and their progress milestones in the continuum of their professional development as budding dental professionals.

Author contributions

AR: Conceptualization, Project administration, Resources, Writing—original draft, Writing—review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

The author would like to thank the Clinical Education team at the University of Edinburgh for their constructive feedback and support to the coursework, which made this article possible.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Komabayashi T, Razak AAA, Bird WF. Dental education in Malaysia. *Int Dent J.* (2007) 57:429–32. doi: 10.1111/j.1875-595X.2007.tb00145.x

2. M.Q.A. *Code of Practice for Programme Accreditation – Undergraduate Dental Degree_version 2.* Malaysia: Malaysian Dental Council (2019).

3. M.D.C. *Competencies of New Dental Graduates.* Malaysia: Malysian Dental Council (2013).

4. Albino JEN, Young SK, Neumann LM, Kramer GA, Andrieu SC, Henson L, et al. Assessing dental students' competence: best practice recommendations in the performance assessment literature and investigation of current practices in predoctoral dental education. *J Dent Educ.* (2008) 72:1405–35. doi: 10.1002/j.0022-0337.2008.72.12.tb04620.x

5. Khanna R, Mehrotra D. The roadmap for quality improvement from traditional through competency based (CBE) towards outcome based education (OBE) in dentistry. *J Oral Biol Craniofac Res.* (2019) 9:139–42. doi: 10.1016/j.jobcr.2019.02.004

6. Chuenjitwongsa S, Oliver RG, Bullock AD. Competence, competency-based education, and undergraduate dental education: a discussion paper. *Eur J Dent Educ.* (2018) 22:1–8. doi: 10.1111/eje.12213

7. Bloom BS, Engelhart MD, Furst ER, Hill WR, Kratwohl DR. *Taxonomy of Educational Objectives: The classifIcation of Educational Goals. Handbook I, Cognitive Domain.* New York, NY: Longmans Green (1956).

8. Miller GE. The assessment of clinical skills/competence/performance. *Acad Med.* (1990) 65:S63–7. doi: 10.1097/00001888-199009000-00045

9. Douthit NT, Norcini J, Mazuz K, Alkan M, Feuerstein M-T, Clarfield AM, et al. Assessment of global health education: the role of multiple-choice questions. *Front Public Health.* (2021) 9:640204. doi: 10.3389/fpubh.2021.640204

10. Abouelkheir H. The criteria and analysis of multiple-choice questions in undergraduate dental examinations. *J Dent Res Rev.* (2018) 5:59–64. doi: 10.4103/jdrr.jdrr_30_18

11. Kulasegaram K, Rangachari PK. Beyond "formative": assessments to enrich student learning. *Adv Physiol Educ.* (2018) 42:5–14.

12. Shumway JM, Harden RM. AMEE Guide No. 25: the assessment of learning outcomes for the competent and reflective physician. *Med Teach.* (2003) 25:569–84. doi: 10.1080/0142159032000151907

13. Tavakol M, Dennick R. The foundations of measurement and assessment in medical education. *Med Teach.* (2017) 39:1010–5. doi: 10.1080/0142159X.2017.1359521

14. Anonymous Dental Act. *Laws of Malaysia. ACT 804.* Malaysia: Putrajaya (2018).

15. Boud D, Soler R. Sustainable assessment revisited. *Assess Eval High Educ.* (2016) 41:400–13. doi: 10.1080/02602938.2015.1018133

16. Epstein RM. Medical education - assessment in medical education. *N Engl J Med.* (2007) 356:387–96. doi: 10.1056/NEJMra054784

17. Kibble JD. Best practices in summative assessment. *Adv Physiol Educ.* (2017) 41:110–9. doi: 10.1152/advan.00116.2016

18. Jolly B, Dalton MJ. *Written Assessment.* Chichester, UK: John Wiley and Sons, Ltd (2018).

19. Van Der Vleuten CPM, Schuwirth LWT, Driessen EW, Dijkstra J, Tigelaar D, Baartman LKJ, et al. A model for programmatic assessment fit for purpose. *Med Teach.* (2012) 34:205–14. doi: 10.3109/0142159X.2012.652239

20. Van Der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ Theory Pract.* (1996) 1:41–67. doi: 10.1007/BF00596229

21. Schuwirth LWT, Van Der Vleuten CPM. How 'testing' has become 'programmatic assessment for learning'. *Health Prof Educ.* (2019) 5:177–84. doi: 10.1016/j.hpe.2018.06.005

22. Downing SM. Reliability: on the reproducibility of assessment data. *Med Educ.* (2004) 38:1006–12. doi: 10.1111/j.1365-2929.2004.01932.x

23. Hift RJ. Should essays and other open-ended -type questions retain a place in written summative assessment in clinical medicine? *BMC Med Educ.* (2014) 14:249. doi: 10.1186/s12909-014-0249-2

24. Mirbahai L, Adie JW. Applying the utility index to review single best answer questions in medical education assessment. *Arch Epid Public Health.* (2020) 1:1–5. doi: 10.15761/AEPH.1000113

25. Gerhard-Szep S, Güntsch A, Pospiech P, Söhnel A, Scheutzel P, Wassmann T, et al. Assessment formats in dental medicine: an overview. *GMS Z Med Ausbild.* (2016) 33:Doc65. doi: 10.3205/zma001064

26. Schuwirth LWT, van der Vleuten CPM. How to design a useful test: the principles of assessment. In: Swanwick T, Forrest K, and O'Brien BC, editors. *Understanding Medical Education: Evidence, Theory, and Practice*. Hoboken, NJ: Wiley (2018), p. 277–90. doi: 10.1002/9781119373780.ch20

27. Norcini JJ, Swanson DB, Grosso LJ, Webster GD. Reliability, validity and efficiency of multiple choice question and patient management problem item formats in assessment of clinical competence. *Med Educ.* (1985) 19:238–47. doi: 10.1111/j.1365-2923.1985.tb01314.x

28. Abdul Rahim AF, Simok AA, Abdull Wahab SF. A guide for writing single best answer questions to assess higher-order thinking skills based on learning outcomes. *Educ Med J.* (2022) 14:111–24. doi: 10.21315/eimj2022.14.2.9

29. Schuwirth L, Colliver J, Gruppen L, Kreiter C, Mennin S, Onishi H, et al. Research in assessment: consensus statement and recommendations from the Ottawa 2010 Conference. *Med Teach.* (2011) 33:224–33. doi: 10.3109/0142159X.2011.551558

30. Towns MH. Guide to developing high-quality, reliable, and valid multiple-choice assessments. *J Chem Educ.* (2014) 91:1426–31. doi: 10.1021/ed500076x

31. Case SM, Swanson D. *Constructing Written Test Questions for the Basic and Clinical Sciences*. Philadelphia, PA: National Board of Medical Examiners. (2002).

32. Tarrant M, Ware J, Mohammed AM. An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC Med Educ.* (2009) 9:40. doi: 10.1186/1472-6920-9-40

33. Rodriguez MC. Three options are optimal for multiple-choice items: a meta-analysis of 80 years of research. *Educ Meas Issues Pract.* (2005) 24:3–13. doi: 10.1111/j.1745-3992.2005.00006.x

34. Zoanetti N, Beaves M, Griffin P, Wallace EM. Fixed or mixed: a comparison of three, four and mixed-option multiple-choice tests in a Fetal Surveillance Education Program. *BMC Med Educ.* (2013) 13:35. doi: 10.1186/1472-6920-13-35

35. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. *Appl Meas Educ.* (2002) 15:309–33. doi: 10.1207/S15324818AME1503_5

36. Dellinges MA, Curtis DA. Will a short training session improve multiple-choice item-writing quality by dental school faculty? A pilot study. *J Dent Educ.* (2017) 81:948–55. doi: 10.21815/JDE.017.047

37. Van Der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. *Med Educ.* (2005) 39:309–17. doi: 10.1111/j.1365-2929.2005.02094.x

38. Anonymous. *Standards for Educational and Psychological Testing / American Educational Research Association, American Psychological Association, National Council on Measurement in Education*. Washington, DC: American Educational Research Association (2014).

39. Williams JC, Baillie S, Rhind SM, Warman S, Sandy J, Ireland A. *A Guide to Assessment in Dental Education*. Bristol: University of Bristol (2015).

40. Biggs J. Enhancing teaching through constructive alignment. *High Educ.* (1996) 32:347–64. doi: 10.1007/BF00138871

41. Hamdy H. Blueprinting for the assessment of health care professionals. *Clin Teach.* (2006) 3:175–9. doi: 10.1111/j.1743-498X.2006.00101.x

42. Coderre S, Woloschuk W, Mclaughlin K. Twelve tips for blueprinting. *Med Teach.* (2009) 31:322–4. doi: 10.1080/01421590802225770

43. Biggs J. Constructive alignment in university teaching. *HERDSA Rev High Educ.* (2014) 1:5–22.

44. Sam AH, Hameed S, Harris J, Meeran K. Validity of very short answer versus single best answer questions for undergraduate assessment. *BMC Med Educ.* (2016) 16:266. doi: 10.1186/s12909-016-0793-z

45. Bird JB, Olvet DM, Willey JM, Brenner J. Patients don't come with multiple choice options: essay-based assessment in UME. *Med Educ Online.* (2019) 24:1649959. doi: 10.1080/10872981.2019.1649959

46. Surry LT, Torre D, Durning SJ. Exploring examinee behaviours as validity evidence for multiple-choice question examinations. *Med Educ.* (2017) 51:1075–85. doi: 10.1111/medu.13367

47. Schuwirth LWT, van der Vleuten CPM. Different written assessment methods: what can be said about their strengths and weaknesses? *Med Educ.* (2004) 38:974–9. doi: 10.1111/j.1365-2929.2004.01916.x

48. Abdulghani HM, Irshad M, Haque S, Ahmad T, Sattar K, Khalil MS. Effectiveness of longitudinal faculty development programs on MCQs items writing skills: a follow-up study. *PLoS ONE.* (2017) 12:e0185895. doi: 10.1371/journal.pone.0185895

49. Palmer EJ, Devitt PG. Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? Research paper. *BMC Med Educ.* (2007) 7:49. doi: 10.1186/1472-6920-7-49

50. Capan Melser M, Steiner-Hofbauer V, Lilaj B, Agis H, Knaus A, Holzinger A. Knowledge, application and how about competence? Qualitative assessment of multiple-choice questions for dental students. *Med Educ Online.* (2020) 25:1714199. doi: 10.1080/10872981.2020.1714199

51. Smith CD, Worsfold K, Davies L, Fisher R, McPhail R. Assessment literacy and student learning: the case for explicitly developing students 'assessment literacy'. *Assess Eval High Educ.* (2013) 38:44–60. doi: 10.1080/02602938.2011.598636

52. Patel US, Tonni I, Gadbury-Amyot C, Vleuten CPMVD, Escudier M. Assessment in a global context: an international perspective on dental education. *Eur J Dent Educ.* (2018) 22:21–7. doi: 10.1111/eje.12343

53. Ebel RL. Writing the test item. In: Linquist EF, editor. *Educational Measurement*, 1st ed. Washington, DC: American Council on Education (1951), p. 621–94.

54. Cilliers FJ, Schuwirth LWT, Herman N, Adendorff HJ, Van Der Vleuten CPM. A model of the pre-assessment learning effects of summative assessment in medical education. *Adv Health Sci Educ Theory Pract.* (2012) 17:39–53. doi: 10.1007/s10459-011-9292-5

55. Wallach PM, Crespo LM, Holtzman KZ, Galbraith RM, Swanson DB. Use of a committee review process to improve the quality of course examinations. *Adv Health Sci Educ Theory Pract.* (2006) 11:61–8. doi: 10.1007/s10459-004-7515-8

56. Tarrant M, Ware J. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Med Educ.* (2008) 42:198–206. doi: 10.1111/j.1365-2923.2007.02957.x

57. Barman A. Standard setting in student assessment: is a defensible method yet to come? *Ann Acad Med Singapore.* (2008) 37:957–63. doi: 10.47102/annals-acadmedsg.V37N11p957

58. De Champlain AF. Standard setting methods in medical education. In: Swanwick T, Forrest K, and O'Brien BC, editors. *Understanding Medical Education: Evidence, Theory, and Practice*. Hoboken, NJ: Wiley (2018), p. 347–60. doi: 10.1002/9781119373780.ch24

59. Kane MT. Validation. In: Brennan R, editor. *EducationalMeasurement*, 4th ed. Westport, CT: Praeger (2006), p. 17-64.

60. Schuwirth LWT, van der Vleuten CPM. Programmatic assessment and Kane's validity perspective. *Med Educ.* (2012) 46:38–48. doi: 10.1111/j.1365-2923.2011.04098.x

61. Norcini JJ. Setting standards on educational tests. *Med Educ.* (2003) 37:464–9. doi: 10.1046/j.1365-2923.2003.01495.x

62. Ben-David MF. AMEE Guide No. 18: standard setting in student assessment. *Med Teach.* (2000) 22:120–30. doi: 10.1080/01421590078526

63. Schuwirth LWT, Van Der Vleuten CPM. General overview of the theories used in assessment: AMEE Guide No. 57. *Med Teach.* (2011) 33:783–97. doi: 10.3109/0142159X.2011.611022

64. Bandaranayake RC. Setting and maintaining standards in multiple choice examinations: AMEE Guide No. 37. *Med Teach.* (2008) 30:836–45. doi: 10.1080/01421590802402247

65. Puryer J, O'sullivan D. An introduction to standard setting methods in dentistry. *Br Dent J.* (2015) 219:355–8. doi: 10.1038/sj.bdj.2015.755

66. Tavakol M, Dennick R. Postexamination analysis: a means of improving the exam cycle. *Acad Med.* (2016) 91:1324. doi: 10.1097/ACM.0000000000001220

67. Mcmanus I, Ludka K. Resitting a high-stakes postgraduate medical examination on multiple occasions: nonlinear multilevel modelling of performance in the MRCP(UK) examinations. *BMC Med.* (2012) 10:60. doi: 10.1186/1741-7015-10-60

68. Norcini J, Anderson B, Bollela V, Burch V, Costa MJ, Duvivier R, et al. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. *Med Teach.* (2011) 33:206–14. doi: 10.3109/0142159X.2011.551559

69. Harrison CJ, Konings KD, Schuwirth L, Wass V, Vleuten CPMVD. Barriers to the uptake and use of feedback in the context of summative assessment. *Adv Health Sci Educ Theory Pract.* (2015) 20:229–45. doi: 10.1007/s10459-014-9524-6