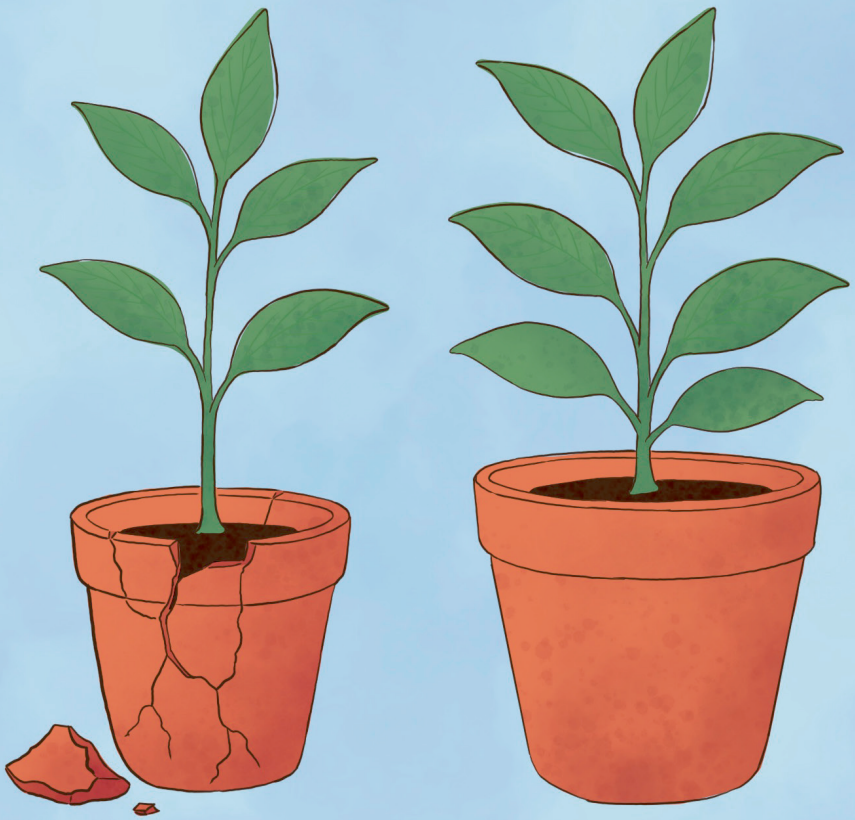
Assessment for growth: fostering student learning through assessment innovations in medical education

Wijk, E.V. van



Chapter 1

General introduction

Introduction

Assessment plays a crucial role in medical education, serving not only as a means to measure knowledge but also as a powerful driver of student learning [1]. Historically, formal assessment began with written and oral examinations [2]. Today, written examinations—the focus of this thesis—are widely used in medical curricula because of their psychometric robustness in evaluating students' knowledge. These assessments primarily consist of multiple-choice questions (MCQs), and are most commonly used as summative assessments to measure student performance through grading [3].

In recent years, the role of assessment has evolved beyond its traditional function of measuring knowledge (assessment *of* learning) toward a more dynamic approach that fosters and facilitates learning (assessment *for* learning) [3, 4]. To align with this shift, traditional assessment tools are being redesigned to enhance in-depth learning and self-regulated learning (SRL) through continuous feedback [5-11]. Emerging technological advancements, such as digital testing and advanced psychometric data analytics, present new opportunities to improve existing assessment methods and develop innovative approaches that emphasize learning. However, the integration of these advancements into medical education and the complex relationship between assessment and learning warrant further exploration.

This thesis investigates innovations in three key aspects of assessment: question design, assessment format, and post-assessment feedback. First, we will investigate the potential of a novel question format: the very short answer question (VSAQ). Secondly, the implementation and benefits of computer adaptive progress testing are researched. Finally, we analyze students' feedback use in formative versus summative assessments and the factors influencing their feedback behaviour. Before discussing these innovations, we will first outline the criteria for good assessment, which are fundamental to improving assessment practices and served as a guiding framework for our studies. This chapter concludes with the research aim of this thesis and an overview of the research projects.

Criteria for good assessment

With the shift in perception that assessment plays a broader role in the learning process by actively stimulating and enhancing learning, it becomes essential to establish criteria that reflect this function. *Table 1* outlines key criteria for effective assessment, along with their definition and educational significance. These criteria are grounded in the framework proposed by Norcini *et al.* and further supported by additional research studies [2, 12-15].

Table 1. Criteria to guide the development of assessments.

Criterion	Definition	Significance
Acceptability	The assessment is fair, reasonable and appropriate, as perceived by the stakeholders (i.e., students, educators and administrators)	Support of the assessment and its results by its users
Authenticity	The assessment challenges students to apply the knowledge and skills they would use in real-life professional scenarios	Encourage students to develop practical and relevant knowledge, preparing them for the challenges they will encounter in their future career
Catalytic effect	The assessment provides feedback that stimulates further learning	Encouragement of learners to reflect on their performance and identify growth opportunities; Driving ongoing learning and development
Cueing effect	The question design unintentionally provides hints that allow students to answer correctly based on recognition rather than true understanding	Ensures that assessments measure genuine knowledge, avoiding inflated scores that do not reflect actual competencies.
Educational effect	The assessment promotes students' preparation beneficial for their learning	Guidance of students towards better understanding of the content
Equivalence	The assessment produces comparable results when administered across different institutions	Fair assessments which measure the same content to the same standard in different contexts
Feasibility	The assessment is practical and realistic to design, implement and score	Effective assessment administration considering the resources, time and effort
Reliability	The assessment produces consistent results if repeated under similar conditions	Consistent and fair assessments which accurately evaluate students' performance
Testing effect	The active retrieval of information from memory during an assessment enhances later recall	Improvement of long-term retention of knowledge
Validity	The assessment accurately measures what it is intended to measure	Reflection of actual competencies or knowledge areas it aims to evaluate

The relative importance of these criteria depends on the assessment's goal and context [2]. For example, summative assessments, which evaluate medical students' knowledge at the end of a course, prioritize psychometric criteria such as ***validity*** and ***reliability*** to ensure the credibility of scores and their implications for academic progression. In contrast, formative assessments – designed to support learning without direct academic consequences – emphasize criteria like the ***catalytic effect*** and learner ***acceptability***. Similarly, different stakeholders, including teachers, learners, and faculty, may prioritize different criteria based on their roles in the assessment process. Learners may value ***reliability*** and ***authenticity***, whereas teachers may focus on ***feasibility*** and ***educational effect***.

These criteria were used to examine the various aspects of assessment in this thesis. While their relative importance varies across the different innovations studied, each innovation aligns with key principles that promote assessment *for* learning, such as the ***catalytic*** and ***educational effect*** (*Table 1*).

Research context

The medical curriculum at Leiden University Medical Center (LUMC) in the Netherlands is structured six-year program divided into a three-year preclinical Bachelor phase and a three-year clinical Master phase. The Bachelor phase includes various theoretical courses. During these courses students are offered the opportunity to practice with the content material through formative assessments. The courses are concluded with written summative assessments, which are graded and contribute to study credits. These assessments mainly consist of MCQs, but other question formats are also used, such as extended matching questions (EMQs), comprehensive integrated puzzles (CIP), or six-step questions. Open essay questions are rarely used due to the extensive time required for grading large student cohorts. After each summative assessment, students receive a form of the Automated Education Evaluation System (AEES)

to evaluate the course. All assessments are conducted digitally in RemindoToets (Paragin). The Master phase focuses on clinical rotations, which are evaluated through a pass/fail decision based on supervisor feedback.

Integral to the medical program is the longitudinal national medical progress test (PT), which medical students of all eight medical schools in the Netherlands are required to take four times annually, resulting in 24 test moments throughout their studies [16]. The PT is a comprehensive written MCQ-test which assesses the end objectives of the medical curriculum across all relevant medical disciplines. The test is designed to monitor students' knowledge progression over time. After each PT, students receive their scores on each discipline and category through an online feedback system (ProF), allowing them to compare their performance with their peers and reflect on their progress with their tutors if desired [17]. In ProF, their (longitudinal) results are both presented in a table and visualized in a graph. Using a longitudinal design and feedback, the PT has an important formative function, amplifying the catalytic effect by promoting ongoing learning and development throughout the medical curriculum [18-22]. The results of the PT are also used for a summative pass/fail decision at the end of each academic year for the awarding of study credits.

The very short answer question

In medical education, multiple-choice questions (MCQs) are widely used due to their ***reliability*** and ease of machine-marking [23, 24]. However, MCQs are often criticized for enabling ***cueing***, where students rely on hints within the questions rather than demonstrating true understanding. This recognition-based study approach does not align well with the educational goals in medical education, which emphasize applying knowledge in complex real-world clinical scenarios [25-27]. Over-reliance on MCQs may therefore limit students' preparedness for such situations [13, 28]. While open-ended essay questions avoid cueing and promote deeper understanding, they are resource-intensive to grade, particularly in large student cohorts. Alternative formats like extended matching questions (EMQ) reduce cueing but still tend to foster recognition-based learning strategies rather than active retrieval [13, 29, 30]. This highlights the need for innovations in question design that support learning.

Very short answer questions (VSAQs)—open-ended questions with responses limited to 1-4 words—offer a promising compromise between MCQs and open essay questions by addressing many limitations of each format. By requiring students to generate answers independently, VSAQs eliminate ***cueing*** and guessing, encourage deeper analytical reasoning and critical thinking (***educational effect*** in *Table 1*) [27, 31-33]. The generation process strengthens memory and supports long-term knowledge retention, aligning with the retrieval effort hypothesis, which suggests that more challenging retrieval tasks enhance learning outcomes [42, 43]. Unlike essay questions, VSAQs are more practical to grade, making them ***feasible*** for large-scale assessments, while maintaining high ***validity***, ***reliability***, and item discrimination in both formative and summative contexts [14, 27, 32, 34, 35]. They also support the catalytic effect (Table 1) by providing meaningful feedback, as open-ended responses allow teachers to gain more insight in students' reasoning processes and identify specific misconceptions or knowledge gaps — insights that are typically less accessible through MCQs [36, 37]. Moreover, VSAQs offer a more ***authentic*** assessment experience by reflecting the open-ended and complex problem-solving required in clinical practice [12, 14, 28].

Despite these clear advantages, the adoption of VSAQs in medical education remains limited. Challenges such as the lack of automated grading tools and concerns about ***feasibility*** have hindered their widespread implementation [14, 34, 38]. Although VSAQs align with key educational goals—such as promoting active learning, providing actionable feedback, and fostering knowledge application—they have yet to be fully utilized in assessment strategies. Addressing these challenges and exploring the broader potential of VSAQs may provide critical insights into how assessments can better support student learning, and preparedness for clinical practice.

Computer adaptive progress testing

Progress testing (PT) was introduced in the 1970s at Maastricht medical school to support problem-based learning (PBL) and discourage test-directed studying by providing longitudinal feedback on students' knowledge progression [17-22]. However, the fixed linear format of the PT, comprised of 200 MCQs, has limitations. It does not account for individual differences in knowledge levels, potentially limiting ***reliability***. Additionally, as the number of participating medical schools has grown, logistical and financial challenges have raised concerns about the ***feasibility*** of the conventional PT [2, 22].

Medical PTs worldwide, including the Netherlands, have adopted formula scoring as a scoring method [48-51]. In formula scoring, correct answers receive points, while incorrect answers incur a penalty, encouraging students to respond only when confident in their answers. A question mark option is included, allowing students to skip a question without a penalty [49]. While intended to reduce the influence of guessing and promote metacognitive reflection, research suggests that the question mark option may introduce bias and affect the test's construct ***validity***, as its use is influenced by individual risk-taking tendencies, metacognitive skills, and self-efficacy [48, 50, 52-54]. Risk-averse students may score lower despite comparable knowledge, and gender differences in guessing behavior further challenge the construct ***validity*** of test scores [52, 55, 56, 57].

Computer adaptive testing (CAT) offers a promising alternative to address the challenges of the conventional PT by tailoring question difficulty to individual ability levels in real time [58]. Since each answer informs the selection of the next question, the question mark option is not feasible in CAT. Based on Item Response Theory (IRT), CAT estimates a test-taker's ability (the 'theta') by assigning difficulty and discrimination parameters to each question [59]. This adaptive approach reduces test length while maintaining or even improving ***reliability*** [60-63]. Moreover, research has shown that CAT can improve motivation and engagement [60, 64, 65]. The Online Adaptive International Progress Test (OAIPT) project has demonstrated the ***feasibility*** of a computer adaptive progress test (CA-PT) across multiple European medical programs [61]. Moreover, its flexibility in test administration alleviates logistical constraints, making large-scale implementation more practical than simultaneous examination.

CAT presents an opportunity to enhance PT by improving efficiency, ***reliability***, and ***feasibility*** while eliminating the need for formula scoring. Before nationwide implementation of CA-PT, it is important to evaluate its psychometric properties to ensure it accurately measures student knowledge. Additionally, understanding the implications of removing the question mark option and how different scoring methods influence student performance is essential to maintaining fairness and ***validity*** in assessment. As medical schools explore the potential transition to CA-PT, these insights will be crucial in guiding the implementation and optimizing progress testing in medical education.

Feedback

The shift from assessment *of* learning to assessment *for* learning in medical education has underscored the importance of formative assessments as tools to promote effective learning strategies and enhance future performance [3, 4]. Formative assessments aim to provide actionable feedback that encourages self-regulated learning and supports students in achieving deeper understanding (***catalytic effect*** in *Table 1*) [4-10, 66]. While the educational value of assessment *for* learning is widely recognized, the impact of formative assessments on students' learning behaviour and motivation remains uncertain. This impact may be influenced by how students perceive and prioritize formative versus summative assessments. Summative assessments, which directly affect final grades, tend to motivate students more, shown by increased test-taking efforts [18, 67-70]. In contrast, formative assessments, due to their lack of direct consequences, may result in lower levels of effort, as suggested by the Expectancy-Value Theory (EVT) [69, 71, 72]. According to EVT—a widely applied test-taking motivation framework—motivation is driven by students' expectations of success and the value they assign to a given task [72]. Understanding how students adapt their learning behaviour, such as test preparation and feedback use, to both formative and summative assessments can help educators optimize these assessments to promote deeper learning and foster lifelong learning skills.

Feedback is central to formative assessments, as it supports the development of learning strategies and self-regulation skills [4-10, 66]. In the context of medical education, PTs provide feedback intended to give students insights into their knowledge progression over time and guide student's learning [17]. Effective use of feedback requires students to develop feedback literacy, which includes understanding feedback's purpose, engaging with it emotionally, and implementing appropriate strategies based on the feedback [10, 73, 74]. Winstone *et al.* [75] identified four key psychological processes essential for effective feedback use: awareness of feedback meaning, cognizance of suitable strategies for implementation, agency in executing strategies, and the volition to explore and act upon feedback. The effectiveness of PT feedback, however, is often limited by students' engagement levels and their capacity to interpret and act on upon the feedback (agency) [76, 77]. Research suggests that students infrequently use PT feedback to reflect on and enhance their learning [69]. This significant loss of formative value points to a need for further research into students' feedback behaviour, specifically the factors that inhibit or promote effective use of PT feedback.

Outline of this thesis

Despite the growing body of knowledge regarding assessment innovations, many questions remain unanswered and require further exploration before these innovations can be successfully and widely implemented in medical curricula. In this thesis, we explore the opportunities and challenges of VSAQs in **Part I**, computer adaptive progress testing in **Part II**, and feedback in **Part III** within undergraduate medical education. Our aim is to optimize the assessment of medical students, improve students' learning and preparation for clinical practice and, ultimately, contribute to the development of more competent clinicians. *Table 2* provides an overview of the studies included in this thesis, detailing the research questions, designs, methods and analyses.

Part I of this thesis comprises three quantitative studies and a viewpoint article, all focusing on the comparison between VSAQs and MCQs. The randomized cohort study in **chapter 2** examines the

psychometric properties, acceptability, cueing effects and student experiences. We partly replicated the study of Sam *et al.* [14] to validate their positive results in another educational setting with less experienced teachers and a different student population. In **chapter 3** we use retrospective data of mixed-format examinations consisting of both VSAQs and MCQs from first- and second-year bachelor students to assess whether VSAQs or MCQs more effective at identifying poor and excellent performing students based on their graded point average (GPA). To enhance our understanding of the effectiveness of different question formats in retrieval practice (i.e., testing effect), we compare the effects of VSAQ and MCQ practice tests on knowledge retention in a real-life educational setting in **chapter 4**. In this within-subjects experimental study we use the scores on a final retention test, three weeks after the last practice test, as a measure for knowledge retention. Additionally, we evaluate students' experiences with the practice tests. In the last chapter of this first part (**chapter 5**) we discuss the current use of MCQs in medical assessment and propose VSAQs as a promising alternative with advantageous properties for both teachers and students.

In **Part II**, the focus shifts to computer adaptive testing in the context of the medical PT. The first study, in **chapter 6**, is a multicentre study which evaluates the correlation between test performance on a CA-PT and a conventional PT. Secondly, we assess the feasibility, student motivation, engagement and experiences with the CA-PT. The cross-over set-up allows for a direct comparison between the two test formats in the same cohort of students within an authentic setting, which is a crucial step toward informed and effective nationwide implementation. The study in **chapter 7** explores the relationship between the use of the question mark option in the conventional PT and student performance on the CA-PT using longitudinal retrospective PT data. We also evaluate the longitudinal reliability and convergent validity of the CA-PT by examining the correlation between the two PT formats over time.

Part III consists of a mixed-methods study and qualitative study focusing on feedback behaviour. The mixed-methods study in **chapter 8** investigates the effect of a PT with a summative component (*summative* PT) and a purely formative PT (*formative* PT) on test preparation, test-taking motivation and feedback use after the test. In this study we use the EVT as theoretical framework [72]. We triangulate quantitative questionnaire data, logging data from ProF, and qualitative interview data to provide a comprehensive understanding of medical' students feedback use and test-taking motivation in a formative and summative setting. In **chapter 9** we present a qualitative study exploring the processes and factors that influence students' use of PT feedback. The data analysis is guided by the psychological processes necessary for effective feedback use, as identified by Winstone *et al.* [75]. This study addresses the knowledge gap regarding effective practices and existing barriers in PT feedback utilization, offering valuable insights into the obstacles that hinder students from effectively using PT feedback.

Finally, the general discussion in **chapter 10** integrates the main findings with the current literature, and provides practical implications and recommendations for future research emerging from this thesis.

Table 2. Overview of research aims, and corresponding study design, method and analyses.

	Chapter	Research aim(s)	Design	Research method	Analyses
Part I	2	1. Externally validate positive results of VSAQs regarding reliability, discrimination and acceptability 2. Explore impact of VSAQs on cueing effects 3. Explore students' experiences of VSAQs	Single-centre cross-over study	Formative assessment and student surveys	Descriptive statistics and psychometric analyses (Cronbach's α, Rir-values)
	3	1. Examine the relationship between question format and academic performance 2. Evaluate the ability of VSAQs and MCQs to identify poor and excellent performing students	Retrospective cohort study	Summative assessment results	Psychometric analyses (Rir-values), linear regression, ROC curves
	4	1. Compare the effect of retrieval effect with VSAQ and MCQ practice tests on knowledge retention 2. Evaluate students' experiences with the practice tests	Single-centre within-subjects experimental study	Practice tests, final test and student surveys	Descriptive statistics and 2x2 repeated measures ANOVA
	5	Viewpoint on assessment questions within medical education	Viewpoint article	Viewpoint and implementation tips based on previous research and own experiences	n.a.
Part II	6	1. Evaluate the correlation between test performance on a CA-PT and a conventional PT 2. Assess the feasibility and student experiences of a CA-PT	Multi-centre cross-over study	Progress test results and student surveys	Descriptive statistics, Pearson correlation coefficient, T-test
	7	1. Explore the relationship between question-mark option use in the conventional PT and student performance on the CA-PT 2. Evaluate the correlation between the conventional PT and CA-PT over time	Multi-centre longitudinal retrospective study	Progress test results	Pearson correlation coefficient, linear regression, model-based cluster analysis
Part III	8	Investigate the effect of a progress test with a summative component and a purely formative progress test on (a) test preparation (b) factors that influence test-taking motivation, and use of feedback (c) self-reported and actual feedback use after the test	Mixed-Methods study	Student surveys, logging data and semi-structured interviews	Descriptive statistics, T-test, chi-squared tests, logistic regression, template analysis with *a priori* themes
	9	Explore which processes and factors affect medical students' feedback use within a Dutch progress testing context	Qualitative study	Semi-structured interviews	Template analysis with *a priori* themes

VSAQ = very short answer question; MCQ = multiple-choice question; PT = progress test; CA-PT = computer adaptive progress test.

References

1. Newble DI, Entwistle NJ. Learning styles and approaches: implications for medical education. Medical Education. 1986;20(3):162-75.
2. 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. Medical Teacher. 2011;33(3):206-14.
3. Schuwirth LWT, Van der Vleuten CPM. Programmatic assessment: From assessment of learning to assessment for learning. Medical Teacher. 2011;33(6):478-85.
4. Scott IM. Beyond 'driving': The relationship between assessment, performance and learning. Medical Education. 2020;54(1):54-9.
5. Berkhout JJ, Helmich E, Teunissen PW, van der Vleuten CPM, Jaarsma ADC. Context matters when striving to promote active and lifelong learning in medical education. Medical Education. 2018;52(1):34-44.
6. Black P, Wiliam D. Assessment and Classroom Learning. Assessment in Education: Principles, Policy & Practice. 1998;5(1):7-74. doi: 10.1080/0969595980050102.
7. Brown GTL, Peterson ER, Yao ES. Student conceptions of feedback: Impact on self-regulation, self-efficacy, and academic achievement. British Journal of Educational Psychology. 2016;86(4):606-29.
8. Castro MABE, de Almeida RLM, Lucchetti ALG, Tibiriçá SHC, da Silva Ezequiel O, Lucchetti G. The Use of Feedback in Improving the Knowledge, Attitudes and Skills of Medical Students: a Systematic Review and Meta-analysis of Randomized Controlled Trials. Medical Science Educator. 2021;31(6):2093-104.
9. Koh LC. Refocusing formative feedback to enhance learning in pre-registration nurse education. Nurse Education in Practice. 2008;8(4):223-30.
10. Watling CJ, Ginsburg S. Assessment, feedback and the alchemy of learning. Medical Education. 2019;53(1):76-85.
11. Hattie J, Timperley H. The Power of Feedback. Review of Educational Research. 2007;77(1):81-112.
12. Gulikers JTM, Bastiaens TJ, Kirschner PA. A five-dimensional framework for authentic assessment. Educational Technology Research and Development. 2004;52(3):67-86.
13. Veloski JJ, Rabinowitz HK, Robeson MR, Young PR. Patients don't present with five choices: an alternative to multiple-choice tests in assessing physicians' competence. Academic Medicine: Journal of the Association of American Medical Colleges. 1999;74(5):539-46.
14. Sam AH, Field SM, Collares CF, van der Vleuten CPM et al. Very-short-answer questions: reliability, discrimination and acceptability. Medical Education. 2018;52(4):447-55.
15. Roediger HL, Butler AC. The critical role of retrieval practice in long-term retention. Trends in Cognitive Sciences. 2011;15(1):20-7.
16. Tio RA, Schutte B, Meiboom AA, Greidanus J, Dubois EA, Bremers AJA, et al. The progress test of medicine: the Dutch experience. Perspectives on Medical Education. 2016;5(1):51-5.
17. Van Der Vleuten CPM, Verwijnen GM, Wijnen WHFW. Fifteen years of experience with progress testing in a problem-based learning curriculum. Medical Teacher. 1996;18(2):103-9.
18. Norman G, Neville A, Blake JM, Mueller B. Assessment steers learning down the right road: Impact of progress testing on licensing examination performance. Medical Teacher. 2010;32(6):496-9.
19. Dion V, St-Onge C, Bartman I, Touchie C, Pugh D. Written-Based Progress Testing: A Scoping Review. Academic Medicine. 2022;97(5):747.
20. Karay Y, Schauber SK. A validity argument for progress testing: Examining the relation between growth trajectories obtained by progress tests and national licensing examinations using a latent growth curve approach. Medical Teacher. 2018;40(11):1123-9.
21. Pugh D, Bhanji F, Cole G, Dupre J, Hatala R, Humphrey-Murto S, et al. Do OSCE progress test scores predict performance in a national high-stakes examination? Medical Education. 2016;50(3):351-8.
22. Van Der Vleuten CPM, Van Der Vleuten CPM. The assessment of professional competence: Developments, research and practical implications. Advances in Health Sciences Education 1996 1:1. 1996/01;1(1).
23. Al-Rukban MO. Guidelines for the construction of multiple choice questions tests. Journal of Family & Community Medicine. 2006;13(3):125-33.
24. Schuwirth L, van der Vleuten C. Written Assessment. ABC of Learning and Teaching in Medicine: Wiley-Blackwell; 2017. p. 65-9.
25. Al-Kadri HM, Al-Moamary MS, Roberts C, Van der vleuten CPM. Exploring assessment factors contributing to students' study strategies: Literature review. Medical Teacher. 2012;34(sup1):S42-S50.
26. Larsen DP, Butler AC, Roediger III HL. Test-enhanced learning in medical education. Medical Education. 2008;42(10):959-66.
27. Sam AH, Wilson R, Westacott R, Gurnell M, Melville C, Brown CA. Thinking differently – Students' cognitive processes when answering two different formats of written question. Medical Teacher. 2021;43(11):1278-85.
28. Bird JB, Olvet DM, Willey JM, Brenner J. Patients don't come with multiple choice options: essay-based assessment in UME. Medical Education Online. 2019;24(1):1649959.
29. Fenderson B, Damjanov I, Robeson M, Veloski J,

Rubin E. The virtues of extended matching and uncued tests as alternatives to multiple choice questions. Human pathology. 1997;28(5).

30. Damjanov I. Testing of medical students with open-ended, uncued questions. Human Pathology. 1995;26(4):362-5.
31. Desjardins I, Touchie C, Pugh D, Wood TJ, Humphrey-Murto S. The impact of cueing on written examinations of clinical decision making: a case study. Medical Education. 2014;48(3):255-61.
32. Sam AH, Hameed S, Harris J, Meeran K. Validity of very short answer versus single best answer questions for undergraduate assessment. BMC Medical Education. 2016;16(1):266.
33. Schuwirth LWT, Vleuten CPM, Donkers HHLM. A closer look at cueing effects in multiple-choice questions. Medical Education. 1996;30(1):44-9.
34. Sam AH, Peleva E, Fung CY, Cohen N, Benbow EW, Meeran K. Very Short Answer Questions: A Novel Approach To Summative Assessments In Pathology. Advances in Medical Education and Practice. 2019;10:943-8.
35. Sam AH, Westacott R, Gurnell M, Wilson R, Meeran K, Brown C. Comparing single-best-answer and very-short-answer questions for the assessment of applied medical knowledge in 20 UK medical schools: Cross-sectional study. BMJ Open. 2019;9(9):e032550.
36. Bala L, Westacott RJ, Brown C, Sam AH. Twelve tips for introducing very short answer questions (VSAQs) into your medical curriculum. Medical Teacher. 2022:1-8.
37. Lertsakulbunlue S, Kantiwong A, Lertsakulbunlue S, Kantiwong A. Development and validation of immediate self-feedback very short answer questions for medical students: practical implementation of generalizability theory to estimate reliability in formative examination designs. BMC Medical Education 2024 24:1. 2024;24(1).
38. Newble DI, Baxter A, Elmslie RG. A comparison of multiple-choice tests and free-response tests in examinations of clinical competence. Medical Education. 1979;13(4):263-8.
39. Dunlosky J, KA R, Marsh E, Nathan M, Willingham D. Improving Students' Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology. Psychological Science in the Public Interest. 2013 Jan;14(1).
40. Butler AC, Karpicke JD, Roediger HL. The effect of type and timing of feedback on learning from multiple-choice tests. Journal of Experimental Psychology Applied. 2007;13(4):273-81.
41. Karpicke JD. Retrieval-based learning: A decade of progress.: Academic Press; 2017.
42. Pyc MA, Rawson KA. Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? Journal of Memory and Language. 2009;60(4):437-47.
43. Samuels SJ. Effects of pictures on learning to read, comprehension and attitudes. Review of Educational Research. 1970;40(3):397-407.
44. Butler AC, Roediger HL. Testing improves long-term retention in a simulated classroom setting. European Journal of Cognitive Psychology. 2007;19(4-5):514-27.
45. Gay LR. The comparative effects of multiple-choice versus short-answer tests on retention. Journal of Educational Measurement. 1980;17:45-50.
46. Greving S, Richter T. Practicing retrieval in university teaching: short-answer questions are beneficial, whereas multiple-choice questions are not. Journal of Cognitive Psychology. 2022;34(5):657-74.
47. Lau KY, Ang JYH, Rajalingam P. Very Short Answer Questions in Team-Based Learning: Limited Effect on Peer Elaboration and Memory. Medical Science Educator. 2023;33(1):139-45.
48. 48. Cecilio-Fernandes D, Medema H, Collares CF, Schuwirth L, Cohen-Schotanus J, Tio RA. Comparison of formula and number-right scoring in undergraduate medical training: a Rasch model analysis. BMC Medical Education. 2017;17:192.
49. Lord FM. Formula scoring and number-right scoring. Journal of Educational Measurement. 1975;12(1):7-11.
50. Muijtjens AM, Mameren HV, Hoogenboom RJ, Evers JL, van der Vleuten CP. The effect of a 'don't know' option on test scores: number-right and formula scoring compared. Medical Education. 1999;33(4):267-75.
51. Wrigley W, Van Der Vleuten CP, Freeman A, Muijtjens A. A systemic framework for the progress test: Strengths, constraints and issues: AMEE Guide No. 71. Medical Teacher. 2012;34(9):683-97.
52. Ravesloot CJ, Van der Schaaf MF, Muijtjens AMM, Haaring C, Kruitwagen CLJJ, Beek FJA, et al. The don't know option in progress testing. Advances in Health Sciences Education. 2015;20(5):1325-38.
53. Rowley G, Traub R. Formula scoring, number-right scoring, and test-taking strategy. Journal of Educational Measurement. 1977;14(1).
54. Kubinger K, Wolfsbauer C. On the risk of certain psychotechnological response options in multiple-choice tests: Does a particular personality handicap examinees? European Journal of Psychological Assessment. 2010;26(4).
55. Messick S. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist. 1995;50(9):741-9.
56. Budescu D, Bar-Hillel M. To Guess or Not to Guess: A Decision-Theoretic View of Formula Scoring. Journal of Educational Measurement. 1993;30(4).
57. Muijtjens A, Mameren H, Hoogenboom R, Evers J, van der Vleuten CPM. The effect of a 'don't know' option on test scores: number-right and formula

scoring compared. Medical education. 1999;33(4).

58. Chang H-H. Psychometrics behind Computerized Adaptive Testing. Psychometrika. 2015;80(1):1-20.
59. Downing SM. Item response theory: applications of modern test theory in medical education. Medical Education. 2003;37(8):739-45.
60. Collares CF, Cecilio-Fernandes D. When I say ... computerised adaptive testing. Medical Education. 2019;53(2):115-6.
61. Rice N, Pêgo JM, Collares CF, Kisielewska J, Gale T. The development and implementation of a computer adaptive progress test across European countries. Computers and Education: Artificial Intelligence. 2022;3:100083.
62. Tian JQ, Miao DM, Zhu X, Gong JJ. An Introduction to the Computerized Adaptive Testing. 2007.
63. Şenel S, Kutlu Ö. Comparison of two test methods for VIS: paper-pencil test and CAT. European Journal of Special Needs Education. 2018;33(5):631-45.
64. Martin A, Lazendic G. Computer-Adaptive Testing: Implications for Students' Achievement, Motivation, Engagement, and Subjective Test Experience. Journal of Educational Psychology. 2017;110.
65. Kisielewska J, Millin P, Rice N, Pego JM, Burr S, Nowakowski M, et al. Medical students' perceptions of a novel international adaptive progress test. Education and Information Technologies. 2023.
66. Schuwirth LWT, van der Vleuten CPM. The use of progress testing. Perspectives on Medical Education. 2012;1(1):24-30.
67. Dijksterhuis MGK, Schuwirth LWT, Braat DDM, Scheele F. An exploratory study into the impact and acceptability of formatively used progress testing in postgraduate obstetrics and gynaecology. Perspectives on Medical Education. 2013;2(3):126-41.
68. Heeneman S, Schut S, Donkers J, van der Vleuten C, Muijtjens A. Embedding of the progress test in an assessment program designed according to the principles of programmatic assessment. Medical Teacher. 2017;39(1):44-52.
69. Schüttpelz-Brauns K, Karay Y, Arias J, Gehlhar K, Zupanic M. Comparison of the evaluation of formative assessment at two medical faculties with different conditions of undergraduate training, assessment and feedback. GMS Journal for Medical Education. 2020;37(4).
70. Wade L, Harrison C, Hollands J, Mattick K, Ricketts C, Wass V. Student perceptions of the progress test in two settings and the implications for test deployment. Advances in Health Sciences Education. 2012;17(4):573-83.
71. Barry CL, Horst SJ, Finney SJ, Brown AR, Kopp JP. Do Examinees Have Similar Test-Taking Effort? A High-Stakes Question for Low-Stakes Testing. International Journal of Testing.2010;10(4):342-63.
72. Eccles JS, Wigfield A. From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. Contemporary Educational Psychology. 2020;61:101859.
73. Carless D, Boud D. The development of student feedback literacy: enabling uptake of feedback. Assessment & Evaluation in Higher Education. 2018;43(8):1315-25.
74. Molloy E, Boud D, Henderson M. Developing a learning-centred framework for feedback literacy. Assessment & Evaluation in Higher Education. 2020;45(4):527-40. doi: 10.1080/02602938.2019.1667955.
75. Winstone NE, Nash RA, Rowntree J, Parker M. 'It'd be useful, but I wouldn't use it': barriers to university students' feedback seeking and recipience. Studies in Higher Education. 2017;42(11):2026-41.
76. Given K, Hannigan A, McGrath D. Red, yellow and green: What does it mean? How the progress test informs and supports student progress. Medical Teacher. 2016;38(10):1025-32.
77. Yielder J, Wearn A, Chen Y, Henning MA, Weller J, Lillis S, et al. A qualitative exploration of student perceptions of the impact of progress tests on learning and emotional wellbeing. BMC Medical Education. 2017;17(1):148.