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

Wijk, E.V. van

Assessment for growth

Fostering student learning through assessment innovations in medical education

Elise van Wijk

Assessment for growth

Fostering student learning through assessment innovations in medical education

Elise van Wijk

Colophon

The research described in this thesis was conducted at the Center for Innovation in Medical Education of Leiden University Medical Center, Leiden, the Netherlands.

Financial support for printing of this thesis was kindly provided by Leiden University, the Dutch Association for Medical Education (NVMO), and Paragin.

Author	Elise van Wijk
Cover design	Eva Taylor Parkins
Layout	Elise van Wijk, Eva Taylor Parkins, Isar de Boer
Printing	Ridderprint | www.ridderprint.nl

Assessment for growth

Fostering student learning through assessment innovations in medical education

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op woensdag 19 november 2025
klokke 14.30 uur

door

Elise Vivianne van Wijk
geboren te Delft
in 1994

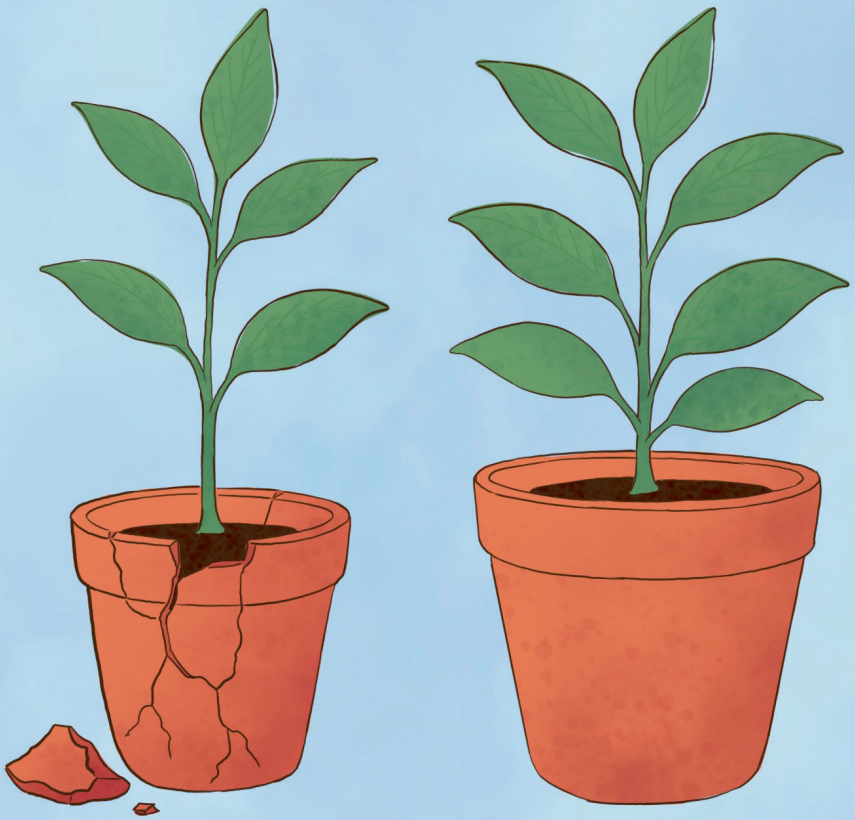
Promotor	Prof. dr. A.M.J. Langers Prof. dr. F.W. Dekker
Copromotor	Dr. F.M. van Blankenstein
Promotiecommissie	Prof. dr. P. Steendijk Prof dr. L.W.T. Schuwirth, Flinders University Prof. dr. M.F. van der Schaaf, University Medical Center Utrecht Prof. dr. S. Heeneman, Maastricht University Dr. M.O. de Jonge, Leiden University Graduate School of Teaching

“What is the fruit of these teachings? Only the most beautiful and proper harvest of the truly educated - tranquility, fearlessness, and freedom. We should not trust the masses who say only the free can be educated, but rather the lovers of wisdom who say that only the educated are free.”

- Epictetus, discourses, 2.1.21-23a

Table of contents

Part III: Feedback



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

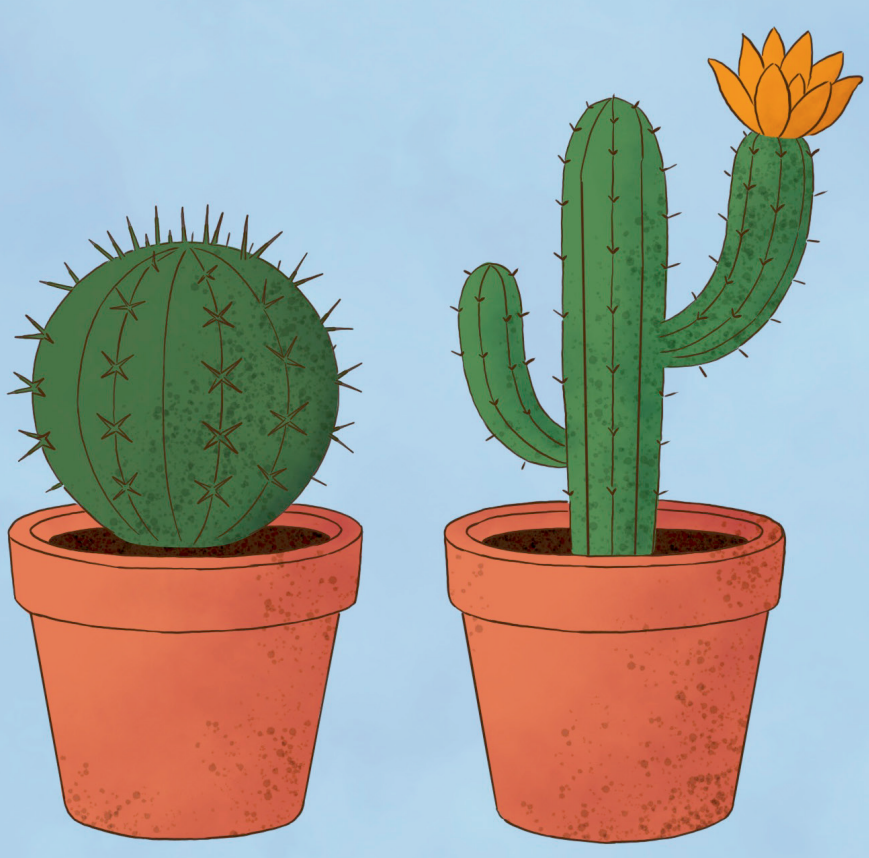
1. Newble DI, Entwistle NJ. Learning styles and approaches: implications for medical education. Medical Education. 1986;20(3):162-75.
2. Norcini J, Anderson B, Bollela V, Burch V, Costa MJ, Duvivier R, et al. Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. 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.



Part I: Very short answer question

Chapter 2

Use of very short answer questions compared to multiple-choice questions in undergraduate medical students: An external validation study

Elise V. van Wijk
Roemer J. Janse
Bastian N. Ruijter
Jos H.T. Rohling
Jolein van der Kraan
Stijn Crobach
Mario de Jonge
Arnout Jan de Beaufort
Friedo W. Dekker
Alexandra M.J. Langers

PLoSONE. 2023;18(7).
DOI: 10.1371/journal.pone.028858

Abstract

Multiple choice questions (MCQs) offer high reliability and easy machine-marking, but allow for cueing and stimulate recognition-based learning. Very short answer questions (VSAQs), which are open-ended questions requiring a very short answer, may circumvent these limitations. Although VSAQ use in medical assessment increases, almost all research on reliability and validity of VSAQs in medical education has been performed by a single research group with extensive experience in the development of VSAQs. Therefore, we aimed to validate previous findings about VSAQ reliability, discrimination, and acceptability in undergraduate medical students and teachers with limited experience in VSAQs development. To validate the results presented in previous studies, we partially replicated a previous study and extended results on student experiences. Dutch undergraduate medical students (*n*=375) were randomized to VSAQs first and MCQs second or vice versa in a formative exam in two courses, to determine reliability, discrimination, and cueing. Acceptability for teachers (i.e., VSAQ review time) was determined in the summative exam. Reliability (Cronbach's α) was 0.74 for VSAQs and 0.57 for MCQs in one course. In the other course, Cronbach's α was 0.87 for VSAQs and 0.83 for MCQs. Discrimination (average Rir) was 0.27 vs. 0.17 and 0.43 vs. 0.39 for VSAQs vs. MCQs, respectively. Reviewing time of one VSAQ for the entire student cohort was ±2 minutes on average. Positive cueing occurred more in MCQs than in VSAQs (20% vs. 4% and 20.8% vs. 8.3% of questions per person in both courses). This study validates the positive results regarding VSAQs reliability, discrimination, and acceptability in undergraduate medical students. Furthermore, we demonstrate that VSAQ use is reliable among teachers with limited experience in writing and marking VSAQs. The short learning curve for teachers, favourable marking time and applicability regardless of the topic suggest that VSAQs might also be valuable beyond medical assessment.

Introduction

Assessment in the educational field commonly uses Multiple Choice Questions (MCQs), because this question type offers high reliability and easy machine-marking. However, it also allows for cueing (i.e., answering questions based on cues in the question or answer options rather than on content knowledge) and stimulates a recognition-based study approach [1–4]. Although recognition may be sufficient to pass a MCQ-based assessment, oftentimes MCQs are not representative for a future situation in which the assessed knowledge has to be applied, for instance because of the absence of a demarcated set of possible answers. This is, among others, the case in medical education, where it has been critically noted that clinical practice does not offer a multiple choice list of possible diagnoses or procedures, nor is there a single best recognisable answer in the medical profession [5, 6].

Although other question formats have been proposed to circumvent the limitations of MCQs such as uncued questions and extended matching questions [6–8], these question formats may still facilitate a recognition-based study approach. Very Short Answer Questions (VSAQs), a free-response type of questions with the answer being limited to 1–4 words, may be better suited to circumvent some of the general limitations of MCQs. The open-ended nature of the VSAQs may prevent surface-level study approaches and cueing [4, 9–11], and it may better represent a profession's real-life practice, such as the medical profession, where VSAQs better reflect clinical practice. In addition, VSAQs are better able to discriminate between students based on proficiency in the content knowledge [9–14] and may increase retention of knowledge [15–17].

Although the use of VSAQs in medical assessments is increasing, evidence regarding validity and reliability of this question type in the medical setting is mainly based on studies from a single research group, consisting of teachers experienced in developing and marking VSAQs [4, 10, 13, 14, 18]. In one study, Sam *et al.* (2018) [13] compared 3rd year medical students starting a test with either VSAQs or MCQs, followed by questions in the opposite format. They observed a higher reliability (Cronbach's α: 0.91 vs. 0.85) and lower mean test score (52.4% vs. 69.7%) for VSAQs vs. MCQs, respectively. Moreover, two-third of students (strongly) agreed that VSAQs better represented clinical practice and half of the students (strongly) agreed VSAQs better prepared them for clinical practice. Cueing was more strongly associated with MCQs. In 5th year pathology students, higher reliability (Cronbach's α: 0.86 vs. 0.76) and a lower median test score (72% vs. 80%) were found in VSAQs vs. MCQs [14]. Lastly, across 20 UK medical schools, Sam and colleagues [18] found a 21% higher test score for MCQs compared to VSAQs, as well as a higher positive cuing rate in MCQs. In this study, they reported marking VSAQs to be feasible.

However, it remains unclear whether the application of VSAQs by teachers with less experience in writing and marking VSAQs, in a different population, country, and medical educational setting yields the same results. Before VSAQs can be implemented in a wider context, more evidence is needed. Therefore, we aimed to externally validate the positive results of VSAQs regarding reliability, discrimination, and acceptability in a cohort of Dutch medical undergraduate students with non-expert teachers. Additionally, we wanted to explore the impact of VSAQs on cueing effects and student experiences of VSAQ-assessment. In order to achieve these aims, we partially replicated the study design of Sam *et al.* [13].

Methods

Setting

This study was simultaneously performed in two different student cohorts (cohort 2019 and cohort 2020) using the same study design. First year students (cohort 2020) followed the fundamental course *'Regulation and Metabolism'* (RM, May 2021) and the second year students (cohort 2019) followed the clinical course *'Diseases of the Abdomen'* (DA, April 2021) in the bachelor of Medicine at the Leiden University Medical Center (LUMC), the Netherlands. Both courses (6 and 7 weeks, respectively) cover metabolic and gastrointestinal topics. During the course, students had weekly mini-exams in DA where they were presented2 –3 VSAQs, but not in RM. Near the end of these courses, students are offered a formative exam and the courses end with a summative exam. After the summative exam, students can evaluate the course with the Automated Education Evaluation System (AEES, *Figure 1A*), which includes questions on constructive alignment. Relevant AEES questions for this study were answered using a 5-point Likert-scale (strongly disagree, disagree, neutral, agree, strongly agree). The LUMC uses RemindoToets (Paragin) [19] for digital assessment with the possibility of proctoring. The current study included the formative and summative exams in both courses for analyses.

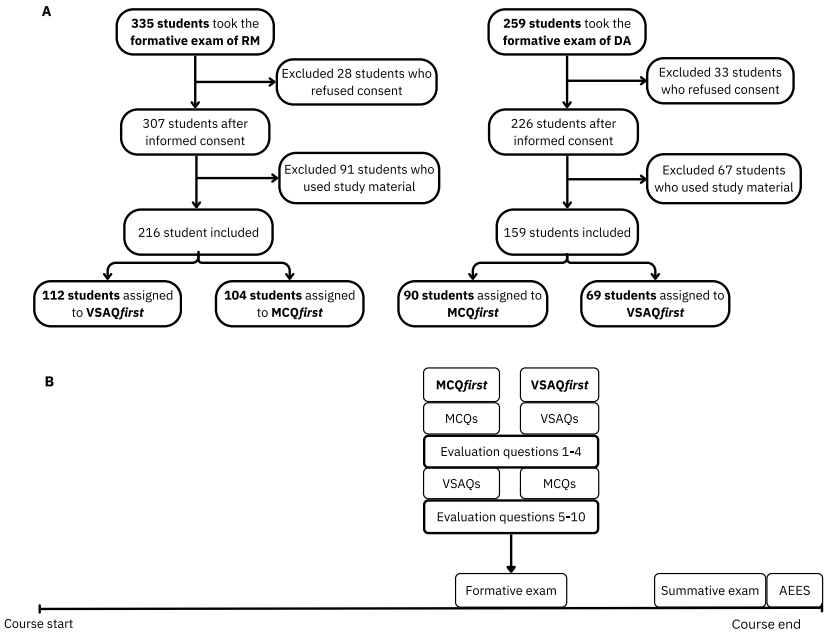


Figure 1. ***(A)*** Set-up of both courses (RM and DA) with the formative exam and contents, summative exam and contents, and the Automated Education Evaluation System (AEES) ***(B)*** Flowchart of the study participants

Formative exam

To determine reliability, discrimination, cueing effects, and students' insights in the formative exams of the RM and DA courses, students were randomly assigned to a group starting with MCQs (RM-MCQ*first* and DA-MCQ*first*) or starting with VSAQs (RM-VSAQ*first* and DA-VSAQ*first*), followed by identical questions in the opposing format, similar to the study of Sam *et al.* [13] (*Figure 1A*). When a section (i.e. either the VSAQ part or the MCQ part) of the exam was finished, students were not able to revisit this specific section. For instance, a student starting the exam with VSAQs could not go back to the VSAQs nor change their answer when they started the section with MCQs. However, it was possible to revisit items or change the response within the same section of the exam before closing the section. The topics in the formative assessment covered the entire spectrum of the course. The MCQs used in the formative exam were written by the course directors with the intent to test the course learning goals. The VSAQs were written in the same way, with assistance from the research team (RJJ, AMJL) to create VSAQs of good quality. The research team was familiar with the literature on VSAQs, but did not yet have any experience in writing VSAQs. For DA, new questions were created as open ended questions suitable for the VSAQ format, and then four answer options for each question were generated to create the parallel MCQ. In RM, existing MCQs that passed the cover test (i.e. the answer can be given without reading the answer options) were transformed into VSAQs by removing the answer options. If the existing MCQs were not specific enough, adjustments to the questions were made to fulfil the VSAQ requirements. Thus, for DA, 24 completely identical questions in both formats were asked, while for RM, 25 questions that tested the same knowledge were asked, albeit sometimes worded differently. The formative exam was available in a fixed timeslot. Participation in this formative exam was mandatory in RM and optional in DA. The exam format order per student was determined using a random number generator in Microsoft Excel (i.e., a Mersenne Twister algorithm) [20]. Only students who gave informed consent were included in the analysis.

After having finished the first part of the formative exam (either MCQs or VSAQs) students were asked to rate three statements on a 5-point Likert scale (strongly disagree, disagree, neutral, agree, strongly agree) and one question ranging from 1 to 10, based on the specific question format with which they were just tested: *1) The questions are a good representation of how I be expected to answer questions in clinical practice; 2) I found the questions easy; 3) I was often unsure whether my answer would be correct); 4) If I had to give an estimate of the grade I would have achieved based on these questions, my estimate would be <grade>.* Because these statements were presented to the students after they finished the first part of the exam, half of the students answered the questions after having finished MCQs only and half of the students after having finished VSAQs only. After the students answered the four evaluation questions, they continued with the second part of the formative test, in which they had to answer the exam-questions in the opposite format. At the end of the second part of the formative exam, all students were asked to rate six more general statements about both question formats: *5) VSAQs are easier than MCQs; 6) VSAQs are more in line with daily clinical practice than MCQs; 7) I prepare differently for an assessment with VSAQs than for an assessment with MCQs; 8) VSAQs would be a better preparation for clinical practice than MCQs; 9) Through the use of VSAQs, the test is better aligned with this course, than a test using MCQs; and 10) Any comments I would like to add: <open question>.* Finally, for research purposes, students were asked whether they used during the formative exam. Given that there were no negative repercussions to using study materials and this was clear to students, we believe that the answer to this question reflects the actual use of study material in the majority of students.

Students who used study materials were excluded from the analysis.

We determined reliability and discriminative capability for content knowledge. The average score, calculated over MCQs and VSAQs separately, was stratified by whether students took MCQs or VSAQs first. Cueing was measured by comparing the answer to an MCQ with the answer to the corresponding VSAQ. We looked at cueing per question (i.e., how often did cueing occur per individual question) and cueing per person (i.e., in how many questions did cueing occur per individual student). We discerned positive and negative cueing. In positive cueing, students used clues in either the MCQ question and/or answer options to arrive at the right answer, which was not possible in VSAQs because no answer options were available. In our study, this could be observed when a student answered a VSAQ incorrectly, but the equivalent MCQ correctly. Negative cueing happens when students are misled by an incorrect answer option in a MCQ (e.g., due to a distractor that is too plausible). In our study, this was derived from a student being able to answer the VSAQ correctly, but not able to give the correct answer to the equivalent MCQ. Although it might have been of influence, the probability of guessing the right answer could not be taken into account. Students' insights were determined from the evaluation questions asked midway through and at the end of the formative exam.

Summative exam

The summative exams of RM and DA were rewritten to replace part of the MCQs with VSAQs (45 in RM and 16 in DA). For RM, this was done through rewriting existing MCQs, whereas for DA, a 2-hour workshop was organized for teachers on how to write VSAQs. Question writers in both courses received written instructions about how to write VSAQs based on information provided by the author of the initial paper on VSAQs, as can currently be found in the publication by Bala *et al.* [13, 21]. For each learning goal, questions were initially written by the experts with content knowledge (e.g., gynaecology questions were written by a gynaecologist). The research team (RJJ, AMJL) assisted, if necessary, in adjusting the questions to the optimal VSAQ format. At the end of both exams, preapproved answers to the VSAQs were automatically marked as correct. Subsequently, teachers reviewed all incorrect answers and could easily add answers that were not in the predefined list, but were also found to be correct. The grading was done by one teacher; a second teacher was consulted when there were doubts about certain answers.

VSAQ review time per question for each teacher was recorded in DA to determine acceptability. The total reviewing time per question was recorded by the reviewer using the timer function on a smartphone. Reviewing time started when the reviewer first looked at the question and ended when the question was fully resolved. This included both reviewing the answers and discussion with other teachers when necessary. Because the marking of only a few VSAQs of the exam was recorded during the initial data collection, which impeded a correct and unselected overview of the reviewing time, one year later the reviewing time of all VSAQs in that year's summative exam was collected again. The AEES questionnaire was supplemented with two questions regarding students' insights: *1) Because I knew that I would be tested by very short answer questions, I studied in another way than I normally would; and 2) Through the use of very short answer questions, the test was a better representation of what I learned in this course, compared to a test using multiple-choice questions.* In addition, the perceived alignment between teaching and assessment was compared to the alignment of the course in the years 2017, 2018, and 2019, determined from two pre-existent questions in the AEES questionnaire: *1) The assessment as*

a whole (form and content) is appropriate for what you should have mastered at the end of the course; and 2) The (online) test formats (e.g. MCQs, open questions, oral and written presentations, practical assessments) matched what I have learned. Due to emergency remote teaching during COVID-19, 2020 is not considered in these comparisons.

Statistical analysis

Continuous variables are presented as mean (standard deviation) or median (interquartile range) depending on their distribution. Categorical variables are presented as number (proportion). Reliability was determined by calculating the Cronbach's α or the VSAQs and MCQs in both formative exam formats, which is a measure of internal correlation between items on a test level [22, 23]. A higher Cronbach's α indicates better reliability with values of 0.7 or higher indicating acceptable reliability. The discriminative capability for content knowledge was determined using the mean of the Rir-values of each question, where the Rir-value is the correlation between one test-item and other test-items [24]. Items with a Rir-value of more than 0.25 typically represent items with an adequate discriminative capability. Mean test scores were calculated as the percentage of correctly answered questions. Reviewing time was expressed in minutes and seconds. All statistical analyses were performed using R version 4.1.0 (R Foundation for Statistical Computing, Vienna, Austria).

Ethical approval

This study was reviewed and approved by the Educational Research Review Board of the Leiden University Medical Center (file number: OEC/ERRB/20201208/1).

Results

Of the 335 students who took the formative exam in RM, 216 students were included in our study. In DA, 159 of the 259 students who took the formative exam were included (*Figure 1B*). In RM, 104 students started with MCQs (RM-MCQ*first*) and 112 students started with VSAQs (RM-VSAQ*first*). In DA, 90 students were assigned to DA-MCQ*first* and 69 students to DA-VSAQfirst The summative exam was made by 352 students in RM and 308 students in DA.

Reliability and discrimination

We compared the VSAQs of students starting with VSAQs with the MCQs of students starting with MCQs. This comparison reflects the results of the VSAQs and MCQs that are not influenced by prior questions. VSAQs had higher reliability compared to MCQs (Cronbach's α 0.74 vs. 0.57 in RM; 0.87 vs. 0.83 in DA for VSAQs vs. MCQs, respectively) (*Table 1*). In the same students, discrimination (mean [SD]), expressed as the Rir-value, was higher in VSAQs compared to MCQs (0.27 [0.15] vs. 0.17 [0.13] in RM; 0.43 [0.10] vs. 0.39 [0.10] in DA, for VSAQs vs. MCQs, respectively). The mean scores (mean [SD]) were lower and had a wider distribution width for VSAQs compared to MCQs (57.0 [15.7] vs. 71.2 [12.2] in RM; 51.6 [23.9] vs. 70.0 [19.7] in DA, for VSAQs vs. MCQs, respectively). These results were similar when comparing results within groups (e.g., VSAQs vs. MCQs within MCQ*first*).

Table 1. Cronbach's alpha, average Rir score and mean (SD) scores for the MCQs and VSAQs in MCQ*first* and VSAQ*first.*

	Regulation and Metabolism				Diseases of the Abdomen			
	MCQ*first*		VSAQ*first*		MCQ*first*		VSAQ*first*	
	MCQ (*n*=104)	VSAQ (*n*=104)	VSAQ (*n*=112)	MCQ (*n*=112)	MCQ (*n*=90)	VSAQ (*n*=85)	VSAQ (*n*=69)	MCQ (*n*=64)
Cronbach's α	0.57	0.61	0.74	0.71	0.83	0.90	0.87	0.71
Average Rir (SD)	0.17 (0.13)	0.19 (0.14)	0.27 (0.15)	0.27 (0.09)	0.39 (0.10)	0.49 (0.13)	0.43 (0.10)	0.26 (0.10)
Mean score (SD), %	71.2 (12.2)	72.3 (12.9)	57.0 (15.7)	75.4 (14.1)	70.0 (19.7)	58.4 (25.9)	51.6 (23.9)	72.5 (15.0)

MCQ = multiple choice question; VSAQ = very short answer question; SD = standard deviation.

Acceptability

In the initially collected data, the average reviewing time per VSAQ by one teacher in the summative exam of DA (7 VSAQs, 308 students) was 2 minutes and 20 seconds (SD 52 seconds). Additionally, on average 2 minutes and 9 seconds (SD 2 minutes and 36 seconds) were spent replying to comments and consultation of other teachers. The maximum time spent on a single VSAQ was 11 minutes and 24 seconds. One year later (22 VSAQs, 338 students), the average time spent on reviewing questions in DA was 1 minute and 58 seconds (SD 40 seconds) and consultation of other teachers took on average 36 seconds (SD 47 seconds).

Secondary outcomes

Positive cueing, defined as a correctly answered MCQ with an incorrectly answered equivalent VSAQ, occurred on average more often per student in RM-VSAQ*first* and DA-VSAQ*first* (20.0%, IQR; 16.0–28.0%; 20.8%, IQR; 12.5–29.2%, respectively) compared to RM-MCQ*first* and DA-MCQ*first* (4.0%, IQR; 4.0–8.0%; 8.3%, IQR; 4.2–16.7%, respectively) (*Table 2*). On a question level, positive cueing occurred in 100% of questions in all groups. The frequency of positive cueing per question was on average higher in RM-VSAQ*first* and DA-VSAQ*first* (14.3%, IQR 7.1–33.9%; 22.7%, IQR 10.9–28.5%, respectively) compared to RM-MCQ*first* and DA-MCQ*first* (4.8%, IQR 2.9–9.6%; 15.9%, IQR 11.8–20.3%, respectively) (*Table 3*). Negative cueing in students, which was defined as students answering the VSAQ correctly and the equivalent MCQ incorrectly, occurred more often per student in RM-MCQ*first* compared to RM-VSAQ*first* (8.0%, IQR 4.0–12.0% to 4.0%, IQR 0.0–4.0%). In DA-MCQ*first*, negative cueing was on average not observed in students (0.0%, IQR 0.0–4.2%). Negative cueing per question occurred in 92%, 56%, 79%, and 79% of the questions for RM-MCQ*first*, RM-VSAQ*first,* DA-MCQ*first*, and DA-VSAQ*first*, respectively. The frequency of negative cueing per question was on average lower in RM-VSAQ*first* compared to RM-MCQ*first* (0.9%, IQR 0.0–1.8%; 3.8%, IQR 1.9–12.5%), but higher in DA-VSAQ*first* compared to DA-MCQ*first* (3.1%, IQR 1.6–5.1%; 1.8%, IQR 1.2–3.5%). The maximum percentage of positive cueing by students in a single question was the highest in RM-MCQ*first* (62.5%). The maximum percentage of negative cueing by students in a single question was 38.5% in RM-MCQ*first*.

Table 2. Positive and negative cueing per person in MCQ*first* and VSAQ*first*.

	Regulation and Metabolism		Diseases of the Abdomen	
	MCQ*first* (*n*=104)	VSAQ*first* (*n*=112)	MCQ*first* (*n*=90)	VSAQ*first* (*n*=69)
Positive cueing, median (IQR), %	4.0 (4.0 - 8.0)	20.0 (16.0 - 28.0)	8.3 (4.2 - 16.7)	20.8 (12.5 - 29.2)
Negative cueing, median (IQR), %	8.0 (4.0 - 12.0)	4.0 (0.0 - 4.0)	0.0 (0.0 - 4.2)	4.2 (0.0 - 4.2)

MCQ = multiple choice question; VSAQ = very short answer question; IQR = interquartile range.

Table 3. Positive and negative cueing per question in MCQ*first* and VSAQ*first*.

	Regulation & Metabolism		Diseases of the Abdomen	
	MCQ*first* (*n*=104)	VSAQ*first* (*n*=112)	MCQfirst (*n*=90)	VSAQ*first* (*n*=69)
Frequency of questions where cueing occurred, %				
Positive cueing	100	100	100	100
Negative cueing	92	56	79	79
Average frequency of cueing per question, %				
Positive cueing, median (IQR)	4.8 (2.9-9.6)	14.3 (7.1 - 33.9)	15.9 (11.8 - 20.3)	22.7 (10.9 - 28.5)
Positive cueing, max	26.9	62.5	32.9	43.8
Negative cueing, median (IQR)	3.8 (1.9-12.5)	0.9 (0.0 - 1.8)	1.8 (1.2 - 3.5)	3.1 (1.6 - 5.1)
Negative cueing, max	38.5	16.1	7.1	10.9

MCQ = multiple choice question; VSAQ = very short answer question; IQR = interquartile range.

When asked whether they found the questions easy, students who had been answering only VSAQs more often disagreed compared to students who had been answering MCQs only in the DA course (EQ2: 3, IQR 2–3 *vs*. 2, IQR 2–2), but students estimated their final grade to be higher if they had started with VSAQs (*Appendix 1 – Supplemental Table 1*). More than 80% of students were uncertain about answering VSAQs correctly (86% and 82% in RM and DA, respectively) (*Appendix 2 – Supplemental Table 2; Figure 2*). Also at the end of the formative exam, after having answered questions in both formats, approximately 90% of students (strongly) disagreed that VSAQs were easier than MCQs (*Appendix 3 – Supplemental Table 3*). 51% of students in RM and 46% in DA (strongly) agreed that assessment with VSAQs changed their test preparation. In DA, 60% of students agreed or strongly agreed that VSAQs better represented clinical practice. This was 34% in RM. Almost 70% of students in RM and 48% in DA (strongly) disagreed that the test was better aligned with the course by using VSAQs. 45% of students in RM and 42% in DA (strongly) agreed they would change learning behavior if tested with VSAQs (*Appendix 4 – Supplemental Table 4*). 83% of students in RM (strongly) disagreed with the statement that the use of VSAQs made the exam a better representation of what they learned during the course compared to MCQs. This was 51% in DA. Perceived alignment of assessment, teaching and learning activities are reported in *Supplemental Table 5* (*Appendix 5*).

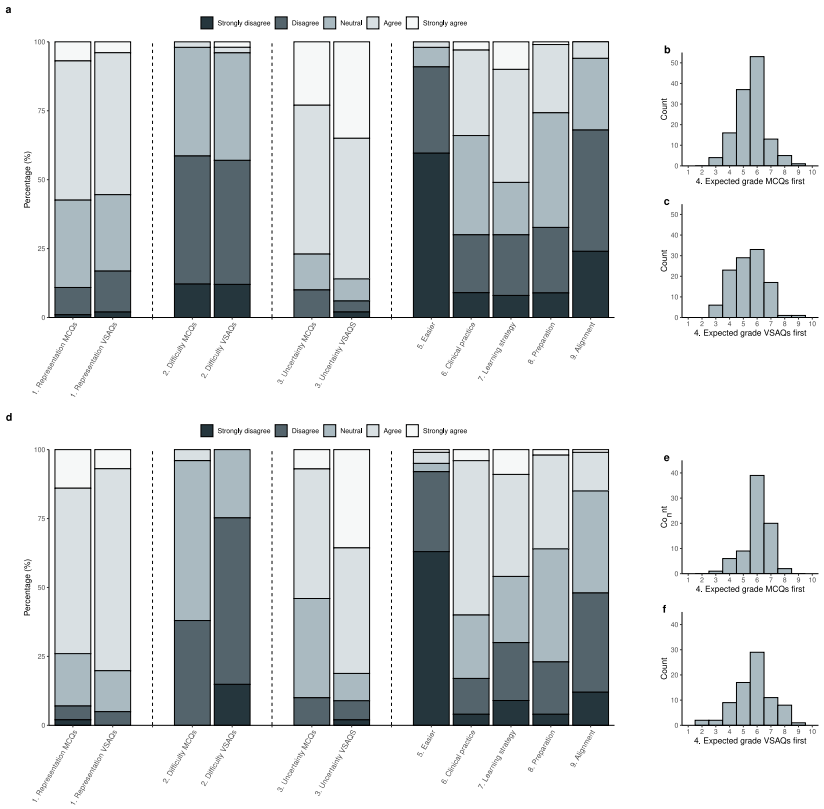


Figure 2. Students' experiences and grade estimates of the MCQs and VSAQs in the formative exam. Distribution of the answers given to the 5-point Likert-scale evaluation questions halfway through the exam after the MCQs or VSAQs and at the end of the exam; and estimates of their grade halfway through the exam in RM ***(A, B, C)*** and DA ***(D, E,F)***.

Discussion

In this study we aimed to externally validate the earlier results regarding reliability, discrimination, and acceptability of VSAQs compared to MCQs in a cohort of Dutch medical undergraduate students, based on earlier work by Sam *et al.* [13]. In accordance with their findings, we observed higher reliability and discrimination of VSAQs compared to MCQs, with an acceptable time to mark VSAQs. Results were more positive in DA than in RM, which might be attributable to the workshop offered to the teachers, better suitable course material, and the opportunity for students to practice with the VSAQs prior to the exams. Additionally, we explored the impact of VSAQs on cueing effects, perceived alignment between assessment and teaching, and student experiences of VSAQs. Cueing effects occurred less frequently in VSAQs compared to MCQs. Students noted a high level of uncertainty when answering VSAQs and around half of students prepared differently for VSAQs. More than half of the students thought VSAQs better represented clinical practice. However, perceived constructive alignment seemed to diminish in RM and not improve in DA.

The higher reliability and discrimination but lower test scores of VSAQs compared to MCQs may in part reflect the decreased possibility of guessing correctly in VSAQs, and are line with Sam *et al.* [13] and other previous studies [14, 18]. The lower score also suggests that VSAQs are more difficult, possibly due to a need of answer generation, rather than answer recognition, which provides a better measure of a students' true content knowledge and increases validity [4, 13, 14]. The high discriminative capability of VSAQs is further supported by higher average Rir values of VSAQs in DA. In RM, average Rir values were relatively low for both MCQs and VSAQs, although an increase in Rir value in VSAQs compared to MCQs could still be observed.

The teachers who graded the VSAQs deemed the reviewing time of VSAQs acceptable. This is supported by previous studies that found comparable and shorter review times, using different marking systems, multiple examiners, and more questions [13, 14]. Nonetheless, whereas not every MCQ has to be reviewed, it should be noted that a VSAQ should always be reviewed after machine marking, although repeated use of questions may decrease reviewing time, depending on software used [13].

Positive cueing per student occurred more often in the students who started with VSAQs, which is in line with the findings of Sam *et al.* [13]. This is expected, as students answering the VSAQs first and MCQs second cannot carry over the MCQ answer to the VSAQ, therefore having to rely on content knowledge for the VSAQ. Cueing per question was also seen more often in this group, but not for every question [13]. However, we most likely also measured students guessing the right answer, as it is nearly impossible to separate guessing and cueing in MCQs [11]. Negative cueing differed only slightly between groups, similar to Sam *et al.* who observed similar negative cueing between groups [13]. It should be noted that in many questions cueing occurred, but per question cueing was observed in few students.

Looking at students' experiences, we found results comparable with Sam *et al.* [13]. The vast majority of the students thought the VSAQs were more difficult than MCQs and almost half of the students said they changed their learning behavior because they were assessed with VSAQs. We observed several noteworthy differences in student experiences between courses that may serve as primer for future research. Concerning clinical practice, students of DA were more positive than students of RM, possibly due to the clinical content in DA having been a better fit for VSAQs than the more fundamental content of RM. This indicates the importance of identifying areas that will benefit most from assessment with VSAQs [21]. Feedback provided by students mainly indicated that VSAQ phrasing might not always have been clear enough. This led to uncertainty regarding the level of specificity of the desired answer, highlighting the importance of a well-designed VSAQ with specific lead-ins [4, 21, 25]. Additionally, a majority of students in RM considered VSAQs to be a poorer representation of course content, while this was only half of the students in DA. This may in part be due to the differences in course content, but uncertainty as a result of an unclearly formulated question may also have played a role. The student feedback in RM possibly also reflects insufficient preparation for the new question format during the course, as students in DA were exposed to VSAQs at multiple timepoints throughout the course. If students have more time to practice, their ability to answer VSAQs may improve [21].

Study strengths are the randomized design, studying two different courses, and the investigation of student perspectives. Furthermore, the fact that teachers who participated in our study had limited experience

with VSAQs allowed us to validate the previous results in an independent setting with less experienced teachers. Limitations are the seemingly poor question quality in the formative RM exam, and the relatively small sample size. Furthermore, due to the low-stakes nature of the formative exam, we cannot be certain that students performed at their best when answering the questions. To determine acceptability, we used only one reviewer who logged the times by hand, leading to less accurate reviewing times. To obtain a more precise measure of acceptability, these findings could be extended by using multiple examiners, more VSAQs and automatically logged times.

Although we validated the VSAQs and investigated student experiences in a medical cohort, we believe that the strengths of VSAQs compared to MCQs are generalizable to other educational fields. Especially, student experiences were mainly related to VSAQs without a focus on a medical context. Real life situations rarely offer a clear single best answer or a list of possible answers. Moreover, in any field open essay questions or other higher-order questions are costly to implement. Although further studies should extend these results to general higher education, our results show VSAQs may provide a promising alternative to MCQ-based assessment in education in general.

Conclusion

In conclusion, this study confirms the positive results of Sam *et al.* [13] on VSAQs in terms of reliability, discrimination, and acceptability in formative assessments in a Dutch cohort of undergraduate medical students. Additionally, these results were confirmed in teachers with only limited prior VSAQ experience and previous results on student experiences are extended. Wider implementation of VSAQs in medical education seems justified and may also improve assessment in other fields of higher education.

References

1. Al-Kadri HM, Al-Moamary MS, Roberts C, Van der Vleuten CP. Exploring assessment factors contributing to students' study strategies: literature review. Med Teach. 2012;34 Suppl 1:S42-50.
2. Eagle M, Leiter E. Recall and Recognition in Intentional and Incidental Learning J Exp Psychol. 1964;68:58-63.
3. Larsen DP, Butler AC, Roediger HL, 3rd. Test-enhanced learning in medical education. Med Educ. 2008;42(10):959-66.10.1111/j.1365-2923.2008.03124.x.
4. 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. Med Teach. 2021;43(11):1278-85.
5. Elstein AS. Beyond multiple-choice questions and essays: the need for a new way to assess clinical competence. Acad Med. 1993;68(4):244-9.
6. 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. Acad Med. 1999;74(5):539-46.
7. Damjanov I, Fenderson BA, Veloski JJ, Rubin E. Testing of medical students with open-ended, uncued questions. Hum Pathol. 1995;26(4):362-5.
8. Fenderson BA, Damjanov I, Robeson MR, Veloski JJ, Rubin E. The virtues of extended matching and uncued tests as alternatives to multiple choice questions. Hum Pathol. 1997;28(5):526-32.
9. 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. Med Educ. 2014;48(3):255-61.
10. 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(1):266.
11. Schuwirth LW, van der Vleuten CP, Donkers HH. A closer look at cueing effects in multiple-choice questions. Med Educ. 1996;30(1):44-9.
12. Newble DI, Baxter A, Elmslie RG. A comparison of multiple-choice tests and free-response tests in examinations of clinical competence. Med Educ. 1979;13(4):263-8.10.
13. Sam AH, Field SM, Collares CF, van der Vleuten CPM, Wass VJ, Melville C, et al. Very-short-answer questions: reliability, discrimination and acceptability. Med Educ. 2018;52(4):447-55.
14. Sam AH, Peleva E, Fung CY, Cohen N, Benbow EW, Meeran K. Very Short Answer Questions: A Novel Approach To Summative Assessments In Pathology. Adv Med Educ Pract. 2019;10:943-8.
15. Neumann J, Simmrodt S, Teichert H, Gergs U, Troussas C. Comparison of Online Tests of Very Short Answer versus Single Best Answers for Medical Students in a Pharmacology Course over One Year. Education Research International. 2021;2021:1-10.
16. McDaniel MA, Roediger HL, McDermott KB. Generalizing test-enhanced learning from the laboratory to the classroom. Psychonomic Bulletin & Review. 2007;14(2):200-6.
17. 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.
18. 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).
19. Toetsanalyse in RemindoToets [Internet]. Paragin; 2018 [Available from: https://www.paragin.nl/update/toetsanalyse-in-remindotoets/.
20. Matsumoto M, Nishimura T. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Trans Model Comput Simul. 1998;8(1):3–30.
21. Bala L, Westacott RJ, Brown C, Sam AH. Twelve tips for introducing very short answer questions (VSAQs) into your medical curriculum. Med Teach. 2022:1-8.
22. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16:297-334.
23. Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ. 2011;2:53-5.
24. De Champlain AF. A primer on classical test theory and item response theory for assessments in medical education. Med Educ. 2010;44(1):109-17.
25. Scheeres K, Agrawal N, Ewen S, Hall I. Transforming MRCPsych theory examinations: digitisation and very short answer questions (VSAQs). BJPsych Bull. 2022;46(1):52-6.

Appendix

Appendix 1 – Supplemental Table 1

Supplemental Table 1. Median (IQR) scores of the 5-point Likert scale evaluation questions (EQ1-3 and EQ5-9) and estimated grade question (EQ4) in the formative exam. EQ1-4 halfway of the exam after the first MCQs (MCQ*first*) or VSAQs (VSAQ*first*): EC5-EQ9 at the end of the exam after both MCQs and VSAQs.

	Regulation and Metabolism		Diseases of the Abdomen	
	MCQ*first* (*n*=104)	VSAQ*first* (*n*=112)	MCQ*first* (*n*=85)	VSAQ*first* (*n*=64)
EQ1	4 (3-4)	4 (3-4)	4 (3-4)	4 (4-4)
EQ2	2 (2-3)	2 (2-3)	3 (2-3)	2 (2-2)
EQ3	4 (4-4)	4 (4-5)	4 (3-4)	4 (4-5)
EQ4	6 (5-6)	6 (6-7)	5 (4-6)	6 (5-6)
EQ5	1 (1-2)		1 (1-2)	
EQ6	3 (2-4)		4 (3-4)	
EQ7	4 (2-4)		3 (2-4)	
EQ8	3 (2-4)		3 (2-4)	
EQ9	2 (2-3)		2 (2-3)	

MCQ = multiple choice question; VSAQ = very short answer question.

EQ1: *The questions are a good representation of how I would be expected to answer questions in clinical practice.*

EQ2: *I found the questions easy.*

EQ3: *I was often unsure whether my answer would be correct.*

EQ4: *If I had to give an estimate of the grade I would have achieved based on these questions, my estimate would be <grade>.*

EQ5: *VSAQs are easier than MCQs.*

EQ6: *VSAQs are more in line with daily clinical practice than MCQs.*

EQ7: *I prepare differently for an assessment with VSAQs than for an assessment with MCQs.*

EQ8: *VSAQs would be a better preparation for clinical practice than MCQs.*

EQ9: *Through the use of VSAQs, the test is better aligned with this course, than a test using MCQs.*

Appendix 2 – Supplemental Table 2

Supplemental Table 2. Distribution of the answers given to the 5-point Likert scale evaluation questions halfway of the formative exam after MCQs (MCQ*first*) of VSAQs (VSAQ*first*).

	Regulation and Metabolism						Diseases of the Abdomen					
	MCQ*first* (*n*=104)			VSAQ*first* (*n*=112)			MCQ*first* (*n*=85)			VSAQ*first* (*n*=64)		
	EQ1	EQ2	EQ3	EQ1	EQ2	EQ3	EQ1	EQ2	EQ3	EQ1	EQ2	EQ3
1: Strongly disagree	1%	12%	0%	2%	12%	2%	2%	0%	0%	0%	15%	2%
2: Disagree	10%	46%	10%	15%	45%	4%	5%	38%	10%	5%	61%	7%
3: Neutral	32%	39%	13%	28%	39%	8%	19%	58%	36%	15%	25%	10%
4: Agree	51%	2%	54%	52%	2%	51%	60%	4%	47%	74%	0%	46%
5: Strongly agree	7%	0%	23%	4%	2%	35%	14%	0%	7%	7%	0%	36%

MCQ = multiple choice question; VSAQ = very short answer question.

EQ1: *The questions are a good representation of how I would be expected to answer questions in clinical practice.*

EQ2: *I found the questions easy.*

EQ3: *I was often unsure whether my answer would be correct.*

Appendix 3 – Supplemental Table 3

Supplemental Table 3. Distribution of the answers given to the 5-point Likert scale evaluation questions at the end of the formative exam.

	Regulation and Metabolism (*n*=216)					Diseases of the Abdomen (*n*=146)				
	EQ5	EQ6	EQ7	EQ8	EQ9	EQ5	EQ6	EQ7	EQ8	EQ9
1: Strongly disagree	59%	9%	8%	9%	24%	63%	4%	9%	4%	12%
2: Disagree	31%	21%	22%	24%	44%	29%	13%	21%	19%	36%
3: Neutral	7%	36%	19%	42%	26%	3%	23%	24%	41%	37%
4: Agree	2%	31%	41%	25%	6%	4%	56%	37%	34%	14%
5: Strongly Agree	0%	3%	10%	1%	0%	0%	4%	9%	2%	1%

EQ5: *VSAQs are easier than MCQs. EQ6: VSAQs are more in line with daily clinical practice than MCQs.*

EQ7: *I prepare differently for an assessment with VSAQs than for an assessment with MCQs.*

EQ8: *VSAQs would be a better preparation for clinical practice than MCQs.*

EQ9: *Through the use of VSAQs, the test is better aligned with this course, than a test using MCQs.*

Appendix 4 – Supplemental Table 4

Supplemental Table 4. Median (IQR) scores and distribution of the answers given to the 5-point Likert scale evaluation questions after the summative exam.

	Regulation and Metabolism		Diseases of the Abdomen	
	Q1 (*n*=147)	Q2 (*n*=148)	Q1 (*n*=85)	Q2 (*n*=85)
1: Strongly disagree	13%	53%	26%	18%
2: Disagree	26%	30%	27%	33%
3: Neutral	17%	12%	6%	19%
4: Agree	35%	5%	38%	26%
5: Strongly Agree	10%	0%	4%	5%
Median score (IQR)	2 (3-4)	1 (1-2)	1 (2-4)	2 (2-4)

IQR = interquartile range.

Q1: *Because I knew that I would be tested by VSAQs, I studied in another way than I normally would.*

Q2: *Through the use of VSAQs, the test was a better representation of what I learned in this course.*

Appendix 5 – Supplemental Table 5

Supplemental Table 5. Median (IQR) scores of the 5-point Likert scale questions on constructive alignment after the summative exam (1: strongly agree, 2: agree, 3: neutral, 4: agree, 5: strongly agree).

	Regulation and Metabolism				Diseases of the Abdomen			
	Q1		Q2		Q1		Q2	
	N	Median (IQR)	N	Median (IQR)	*N*	Median (IQR)	*N*	Median (IQR)
'16/'17	*NA*	*NA*	197	4 (3-4)	*NA*	*NA*	170	4 (3-4)
'17/'18	50	3 (2-4)	50	4 (2-4)	66	4 (3-4)	66	4 (3-4)
'18/'19	63	3 (2-4)	62	3 (2-4)	62	2 (1-3)	62	2 (1-3)
'20/'21	149	2 (1-3)	149	2 (1-3)	127	3 (2-4)	127	3 (2-4)

IQR, interquartile range.

Q1: *The assessment as a whole (form and content) is appropriate for what you should have mastered at the end of the course.*

Q2: *The (online) test formats matched what I have learned; NA = not available.*

Part I: Very short answer question

Chapter 3

Identifying academic success and underperformance: The discriminative power of very short answer questions and multiple-choice questions

Elise V. van Wijk
Floris M. van Blankenstein
B.N. Ruijter
J.H.T. Rohling
J. van der Kraan
F.W. Dekker
Alexandra M.J. Langers

Submitted to Medical Science Educator (2025)

Abstract

Background: Multiple-choice questions (MCQs) are widely used in medical education, but are criticized for cueing and guessing. Very short answer questions (VSAQs), which require students to generate responses independently, may better assess knowledge. While VSAQs demonstrate higher item discrimination within individual exams, their effectiveness in distinguishing academic performance across multiple assessments remains unclear. This study examines whether VSAQs or MCQs more effectively distinguish students of varying performance levels across multiple summative examinations.

Methods: We analyzed retrospective data from six mixed-format examinations with VSAQs and MCQs of three cohorts of first- and second-year medical students. Academic performance was measured using grade point average (GPA) across assessments. Linear regression assessed the relationship of each question format with GPA, while ROC curves and C-statistics evaluated their ability to identify poor and excellent performing students (lowest and highest quintile of GPA).

Results: VSAQs showed higher item discrimination (Rir-values) than MCQs in all exams. VSAQs also had a stronger positive association with GPA compared to MCQs, and higher C-statistics, indicating superior discriminative ability.

Conclusion: VSAQs outperform MCQs in distinguishing academic performance levels across multiple assessments. Their integration into examinations enhances discriminative ability and may facilitate earlier identification of poor and excellent performing students, enabling targeted interventions and support of students.

Introduction

Assessment of undergraduate medical students predominantly relies on multiple-choice questions (MCQs), especially in the single best answer (SBA) format, due to their high reliability and the efficiency of machine marking [1, 2]. However, MCQs have been criticized for their susceptibility to test-taking strategies, such as cueing – where students use cues in the question or answer options to deduce the correct answer without fully applying content knowledge – and guessing [3-6]. These factors introduce noise into MCQ scores, thereby diminishing their ability to accurately reflect students' true understanding [11]. In contrast, the very short answer question (VSAQ), an open-ended question requiring a concise answer, mitigates these issues by eliminating both cueing and guessing [7]. Consequently, VSAQs tend to exhibit higher item discrimination within individual examinations compared to MCQs [8-11], meaning that they can better differentiate between high- and low-performing students on a given test.

Although VSAQs consistently demonstrate superior item discrimination within single examinations, it remains unclear whether they also offer a better means of distinguishing among students with varying academic performance across multiple examinations. Interestingly, higher item discrimination in a single exam does not necessarily translate into a stronger ability to identify students as poor or excellent performers across several examinations. For example, Eijsvogels *et al.* [12] found that while extended matching questions (EMQs) demonstrate superior item discrimination compared to MCQs within examinations [13], they were less effective at identifying poor performing students, despite being better at identifying excellent performing students. Similarly, despite evidence supporting the higher item discrimination of VSAQs compared to MCQs within individual examinations [8-11], their effectiveness in distinguishing overall student academic performance across multiple assessments remains uncertain.

Beyond evaluating the item discrimination of individual assessment questions, it is important to investigate whether certain question formats are better suited for identifying poor and excellent performing students across multiple examinations. This broader perspective can offer insights into the consistency and robustness of these formats in distinguishing students across different performance levels, and formats with higher discriminative power may facilitate the early identification of underperforming students, thereby enabling timely interventions. In this study, we aim to 1) examine the relationship between question format (VSAQs *versus* MCQs) and academic performance; 2) evaluate the ability of VSAQs and MCQs to identify poor and excellent performing students. To address these aims, we first assess the item discrimination of both question formats within each examination, thereby verifying the assumption that VSAQs have superior discriminative ability within examinations [8-11]. We use the VSAQ- and MCQ-scores from two summative mixed-format examinations administered during the first and second year of an undergraduate medical curriculum. Our analysis includes two student populations: 1) all students who participated in the first-year examination, including those who may later leave the program, and 2) nominal students who participated in both the first- and second-year examinations. The first population offers a broader performance range, particularly among lower performing students, while the second population offers more datapoints (i.e., questions) per student, enhancing the reliability of the analysis.

Methods

Setting

This retrospective cohort study was conducted in Leiden University Medical Center (LUMC), the Netherlands. The Dutch medical curriculum comprises a three-year bachelor's program followed by a three-year master's program. Although the bachelor courses are primarily assessed with MCQs, other formats such as extended matching, comprehensive integrated puzzle, open essay, and VSAQs are also included in the written assessments. Additionally, students participate in longitudinal training on various CanMEDS competencies [14] beyond the role of Medical Expert, such as communication skills, leadership, health promotion, and collaboration. To assess the discriminative ability of VSAQs and MCQs we analyzed summative examinations of two medical bachelor's courses: '*Regulation and Metabolism*' (RM), a first-year fundamental course, and '*Diseases of the Abdomen*' (DA), a second-year clinical course. These courses (6 and 7 weeks, respectively) address metabolic and gastrointestinal topics. In our prior study [8], we compared the psychometric properties of MCQs and VSAQs in formative assessments for both courses. Subsequently, the course coordinators added VSAQs to the previously MCQ-only summative exams, resulting in mixed-format examinations. This mixed format has now been used in both courses for three consecutive years (student cohorts 2020-2021; 2021-2022; 2022-2023), resulting in six summative mixed-format examinations available for our analyses. Near the end of the course and prior to the summative assessment, students were given opportunities to practice with the different question types. All assessments were administered digitally through RemindoToets (Paragin) system [15] and yielded study credits.

Participants

The first population, *RM participants*, included all first-year bachelor medical students from the 2020–2021, 2021–2022, and 2022–2023 cohorts who completed the first sitting of the summative RM assessment. The second population, *RM & DA participants*, comprised all second-year bachelor medical students from the same cohorts who completed the first sitting of both the RM and DA summative assessments in consecutive years. Students who attempted less than 75% of all exams taken into consideration for the GPA calculation were excluded from both populations (n=47).

Study design and data collection

We analyzed data from two summative mixed-format examinations from the RM and DA courses, administered between 2021 and 2024. For each assessment, we extracted individual questions, question formats, question scores (coded as 1 for correct and 0 for incorrect), residual item reliability-values (Rir) for each question, and student IDs from the Remindo assessment system. The Rir-value, representing the correlation between one test item and all other items of the test, measures item discrimination by indicating how well each question correlates with overall student performance on that test (after excluding that specific question) [16]. To assess the comparability of question formats within the examinations, we examined the distribution of each question format across the different themes and cognitive process dimensions based on Bloom's revised taxonomy [17]. Bloom's levels were assessed by two researchers of the research team (EvW, FvB). First, a sample of 50 questions, randomly selected from across the examinations, was reviewed together to establish a shared coding approach. Following this, each researcher coded a subset of 30 questions independently, with any discrepancies discussed to reach

consensus (Cohen's kappa = 0.89 before discussion). The remaining questions were then evaluated by the principal investigator (EvW), with consultation from the second researcher (FvB) in cases of uncertainty.

We measured academic performance with the grade point average (GPA), a widely recognized and standardized indicator of academic success [18, 19]. All exam grades from the first and second-year courses were retrieved from the university's administrative system. These numeric grades, combined with the corresponding study credits, were used to calculate students' GPA. The maximum score is 10, and grades of 5.5 or higher are considered sufficient to earn study credits. Failing grades (defined as <5.5) were included to ensure a realistic measurement of academic performance. We evaluated GPA based on two approaches: one using the most recent grade from the last sitting (including retake exams) and another using only the grade from the first sitting. For the first population (*RM participants*) we calculated the GPA based on first-year courses, while for the second population (*RM & DA participants*) the GPA included courses from both the first- and second-year. We included only exam grades. In courses where non-exam components such as presentations, reports, and participation, contributed to the final grade but did not result in a separate grade, these components were excluded. To ensure a fair representation of these courses, we allocated half of the available study credits to the exam grades, reflecting their partial contribution to the overall course assessment. Based on these GPAs, students were categorized by quintiles into '*poor performing*' (first quintile), '*average performing*' (second through fourth quintile), or '*excellent performing*' (fifth quintile) students.

Data analysis

Descriptive statistics were calculated and presented for each examination, including total average scores, average VSAQ and MCQ scores, the distribution of VSAQs and MCQs across the themes, and the distribution of Bloom's cognitive dimensions assessed by VSAQs and MCQs. To compare the different question formats, we calculated separate z-scores for VSAQs and MCQs based on their respective absolute scores. For the *RM participants*, z-scores were derived from their VSAQ and MCQ scores on the RM assessment. For the *RM & DA participants*, z-scores were calculated separately for VSAQs and MCQs using absolute scores from both the RM and DA assessments, ensuring that each student had one VSAQ z-score and one MCQ z-score reflecting their performance across both exams. Item discrimination was determined using the mean of the Rir-values for each question format [16]. The Rir-value for a each question was calculated in relation to the entire exam.

Linear regression analysis was performed with GPA as linear outcome variable and the z-scores of the VSAQ and MCQ as covariates, separately for each cohort and across all cohorts. First, univariate regression analyses were conducted to assess the variance explained by each question format individually (reported as adjusted R2). Next, both formats were included in a multivariate model to account for shared variance and determine their relative predictive value. Student GPAs from the courses of the first year (RM participants) and the first two year (*RM & DA participants*) were used as outcome. Similarly, the same student GPAs were used to create new categorical variables: '*poor performing*' versus '*other*' (average and excellent) and '*excellent performing*' versus '*other*' (average and poor). Here, '*poor performing*' students were defined as those in the lowest quintile, '*average performing*' in the second through fourth quintiles, and '*excellent performing*' in the fifth quintile. Receiver Operating Characteristic (ROC) curves were used to evaluate the ability of MCQs and VSAQs to distinguish between poor and non-poor performing students,

and also between excellent and non-excellent performing students. The Concordance-statistic (C-statistic or area under the ROC curve) was calculated as a measure of discriminatory power, reflecting how well these formats differentiate between performance levels. Significant differences between the C-statics of VSAQs and MCQs were assessed using paired bootstrapping with 95% confidence intervals. All statistical analyses were performed using R version 4.1.0 (R Foundation for Statistical Computing, Vienna, Austria).

Ethical approval

This study received ethical approval from the LUMC Educational Research Review Board (OEC/ERRB/20241008/1). Informed consent was not required, as no new participants were recruited for this study. The data analyzed were retrospective and already available from the assessment system as part of the students' regular curriculum. We collected no additional demographic data, and all data were aggregated to ensure that individual students could not be identified.

Results

Descriptives and item discrimination

The students included in our study for each mixed-format examination, along with the average total, VSAQ, and MCQ scores are presented in *Table 1*. In all mixed-format examinations, the average Rir-values of the VSAQs were consistently higher compared to the MCQs (*Table 1*). The distribution of question formats across the themes was generally even, with no themes exclusively assessed by a single format (*Appendix 1 – Supplemental Table 1*). Across Bloom's taxonomy levels, the distribution within exams was largely balanced, except in RM, where VSAQs appeared more frequently classified as '*Remember*' questions and less as '*Understand*' questions (*Appendix 2 – Supplemental Table 2*).

Table 1. The scores and Rir-values of VSAQs and MCQs within the mixed-format examinations.

	Total examination		Very short answer questions			Multiple-choice questions		
	Max points	Total score[b] mean (SD)	N	Score, mean (SD)	Rir-values, mean (SD)	N	Score, mean (SD)	Rir-values, mean (SD)
RM 2021[a] (*n*=323)	74.50	53.40 (9.19)	42	27.95 (6.36)	0.30 (0.13)	32	24.50 (3.53)	0.21 (0.11)
RM 2022 (*n*=296)	71	45.49 (8.79)	29	15.53 (4.63)	0.28 (0.11)	36	27.29 (4.61)	0.24 (0.10)
RM 2023 (*n*=287)	73	49.96 (9.74)	31	18.97 (5.40)	0.32 (0.12)	38	29.29 (4.79)	0.24 (0.11)
DA 2022 (*n*=282)	91	59.58 (10.35)	22	13.39 (3.41)	0.30 (0.11)	48	33.34 (4.53)	0.18 (0.10)
DA 2023 (*n*=255)	90	64.38 (9.48)	30	20.75 (4.14)	0.26 (0.09)	40	28.50 (3.99)	0.19 (0.11)
DA 2024 (*n*=260)	71	48.55 (8.18)	31	20.62 (4.18)	0.27 (0.13)	20	13.66 (2.64)	0.19 (0.11)

RM = Regulation and Metabolism; DA = Diseases of the Abdomen; N = number of questions.

[a]RM 2021 & DA 2022 = cohort 2020-2021; RM 2022 & DA 2023 = cohort 2021-2022; RM 2023 & DA 2024 = cohort 2022-2023.

[b]Average total score also includes the scores on question formats other than VSAQs and MCQs (comprehensive-integrated-puzzle (CIP) (maximum points: 4), open-ended question without word limitation (maximum points: 8), 6-step pharmacology question (maximum points: 8), hotspot question (maximum points: 1)).

Relationship between question format and academic performance

We conducted a linear regression analysis to examine the effects of VSAQ and MCQ z-scores on GPA. First, we assessed the variance explained by each question format separately using univariate models (R^2 values). Among all *RM participants*, the R^2 values ranged from 0.62 to 0.66 for VSAQ z-score and from 0.44 to 0.57 for MCQ z-score. In all *RM & DA participants*, the R^2 values ranged from 0.68 to 0.71 for VSAQ z-score and from 0.59 to 0.67 for MCQ z-score.

The multiple regression analyses showed a significant positive association between both VSAQ and MCQ z-scores and GPA (*Table 2*). However, for all *RM participants* (*n*=906), the VSAQ z-score had a stronger association with GPA when using first sitting grades (β = 0.66, t(903) = 21.55, p < .001, 95% CI [0.60, 0.72]) than the MCQ z-score (β = 0.37, t(903) = 12.20, p < .001, 95% CI [0.31, 0.42]), with the model explaining 75% of the variance in GPA (F(2, 903) = 942.44, p < .001, adjusted R^2 = .75). Similarly, in all *RM & DA participants* (*n*=797), the VSAQ z-score had a stronger positive association (β = 0.54, t(794) = 22.23, p < .001, 95% CI [0.49, 0.59]) compared to the MCQ z-score (β = 0.37, t(794) = 15.03, p < .001, 95% CI [0.32, 0.41]), with the model accounting for 68% of the variance (F(2, 794) = 1202.38, p < .001, adjusted R^2 = .68). These results were similar when using last sitting grades and for the analyses of the separate cohorts (*Table 2*).

Table 2. Linear regression model parameters with grade point average as outcome based on the first and last sitting grades.

	First sitting grades						Last sitting grades					
	VSAQ z-score			MCQ z-score			VSAQ z-score			MCQ z-score		
Cohort	**ß**	**t**	**95% CI**	**ß**	**t**	**95% CI**	**ß**	**t**	**95% CI**	**ß**	**t**	**95% CI**
RM participants												
2020-2021[a] (*n*=323)	0.68	14.60	0.59–0.77	0.26	5.65	0.17–0.35	0.61	15.08	0.53–0.69	0.21	5.16	0.13–0.29
2021-2022 (*n*=296)	0.62	12.10	0.52–0.72	0.48	9.35	0.38–0.58	0.49	11.32	0.40–0.57	0.39	9.09	0.31–0.47
2022-2023 (*n*=287)	0.67	12.18	0.57–0.78	0.39	7.12	0.29–0.50	0.56	11.42	0.47–0.66	0.33	6.75	0.24–0.43
Total (*n*=906)	0.66	21.55	0.60–0.72	0.37	12.20	0.31–0.43	0.56	21.31	0.51–0.61	0.30	11.66	0.25–0.36
RM & DA participants												
2020-2021 (*n*=282)	0.52	13.44	0.44–0.59	0.35	9.20	0.28–0.43	0.43	12.25	0.36–0.50	0.31	8.65	0.24–0.38
2021-2022 (*n*=255)	0.49	11.86	0.41–0.57	0.40	9.60	0.31–0.48	0.39	10.49	0.31–0.46	0.35	9.59	0.28–0.43
2022-2023 (*n*=260)	0.61	13.69	0.52–0.70	0.36	7.94	0.27–0.44	0.53	13.57	0.45–0.61	0.29	7.52	0.22–0.37
Total (*n*=797)	0.54	22.23	0.49–0.59	0.37	15.03	0.32–0.41	0.45	20.74	0.41–0.49	0.31	14.44	0.27–0.36

RM=Regulation and Metabolism; DA = Diseases of the Abdomen; VSAQ=very short answer question; MCQ=multiple-choice question; CI=confidence interval. SE = standard error. All p-values <0.001.

[a]Cohort 2020-2021 = RM 2021 & DA 2022; Cohort 2021-2022 = RM 2022 & DA 2023; Cohort 2022-2023 = RM 2023 & DA 2024.

Identification of poor and excellent performing students

We calculated the C-statistic to assess the ability of VSAQ and MCQ scores to identify poor performing (lowest GPA quintile) and excellent performing (highest GPA quintile) students. In all *RM participants* (*n*=906), VSAQ z-scores showed higher C-statistics (i.e., greater discriminative ability) for both poor and excellent performing students compared to MCQ z-scores (*Figure 1A*). However, for poor performing students using first sitting grades, the difference between the C-statistics of VSAQ and MCQ was not significant. In all *RM & DA participants* (*n*=797), the C-statistic of the VSAQ z-score remained significantly higher than that of the MCQ z-score for poor performing students. For excellent performing students, there was no significant difference in discriminative ability (*Figure 1B*). The findings across separate cohort analyses followed the same pattern; either the C-statistic for the VSAQ z-score was significantly higher than that of the MCQ z-score, or there was no significant difference between them (*Appendix 3 & 4 – Supplemental Figure 1 & 2*).

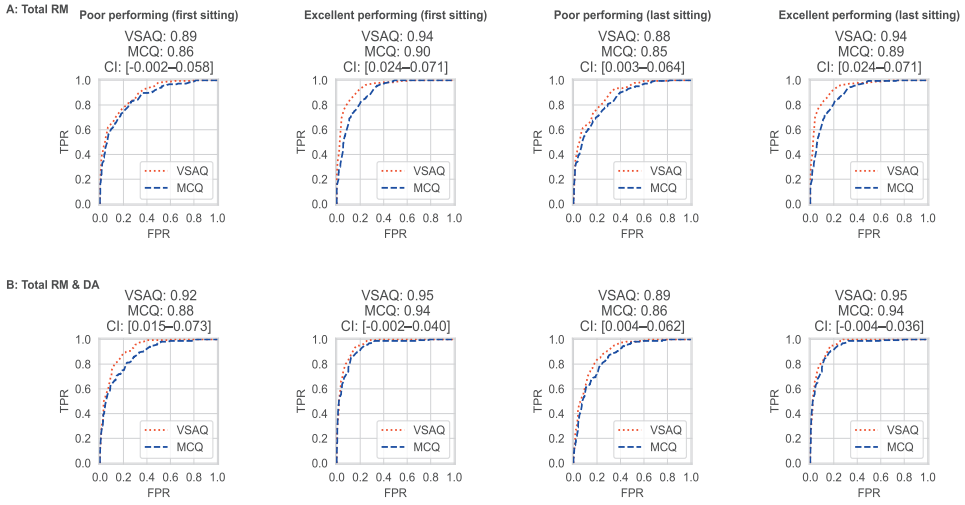


Figure 1. Receiver Operating Characteristic (ROC) curves for both poor and excellent performing students using the first sitting or last sitting grades from A) the total RM participants, and B) the total RM & DA participants. Red = VSAQs; Blue = MCQs; CI = 95% confidence interval; TPR = true positive rate; FPR = false positive rate. C-statistics are shown above each graph, which were calculated from the area under the ROC curve.

Discussion

In this study, we examined whether VSAQs or MCQs are more effective in distinguishing undergraduate medical students with varying academic performance levels based on GPA. Our findings reveal that while both question formats are suitable for distinguishing students across performance levels, VSAQs consistently demonstrate superior effectiveness compared to MCQs in both the linear regression analysis

and ROC curves. Notably, the difference in discriminative ability between VSAQs and MCQs was slightly more pronounced among first-year students (*RM participants*) compared to the combined group of first- and second-year students (*RM & DA participants*). A plausible explanation for this observation is the exclusion of mostly poor performing students who did not nominally progress to the second year, resulting in a more homogeneous group of students with less variability in academic performance. Consistent with previous research [8-11], VSAQs also demonstrated higher item discrimination than MCQs within the individual examinations.

Our findings differ from the findings of Eijsvogels *et al.*, who found that MCQs were more effective than EMQs in identifying poor performing students [12]. This discrepancy may arise from differences in question format and scoring methodology. While EMQs share certain features with VSAQs — such as reduced reliance on guessing compared to MCQs — they still provide a list of plausible options that could lead to guessing. In contrast, VSAQs require students to generate concise responses without external cues, inherently minimizing the likelihood of guessing and offering a more direct measure of their knowledge. Furthermore, Eijsvogels *et al.* [12] penalized incorrect answers on MCQs, potentially increasing the effectiveness of this question format. By normalizing performance scores within mixed-format examinations, we ensured a fair comparison, eliminating potential biases from variations in exam quality or instructional approach. Additionally, our analysis included multiple examinations and a larger sample size, enhancing the robustness of our findings.

The GPA used in this study to assess academic performance is primarily based on assessments that predominantly consist of MCQs, which could bias the results toward MCQs due to their alignment with the format used to calculate GPA. However, despite this potential bias, our findings indicate a clear advantage for VSAQs over MCQs in distinguishing poor and excellent performing students. The superior performance of VSAQs can be attributed to their format, which requires students to generate responses independently, eliminating opportunities for guessing and relying instead on their actual understanding [3-6]. This characteristic makes VSAQs a more accurate reflection of student's knowledge level. While guessing may inflate MCQ scores for poor performing students in a single examination, this effect will diminish when performance is averaged across multiple exams, thereby reducing noise introduced by guessing.

Strengths and limitations

This study is the first to evaluate the discriminative ability of VSAQs across multiple assessments using GPA as a measure of academic performance, rather than focusing exclusively on item discrimination with individual exams. By utilizing real-world summative assessment data from three distinct student cohorts, we minimized the biases associated with voluntary participation and ensured a robust, and representative sample size. This also allowed us to aggregate results and average out variance across examinations and populations, thereby enhancing the reliability of our findings.

The inclusion of both first-year students with greater variation in academic performance, and nominal second-year students allowed for a more comprehensive analysis of performance across different populations, while mitigating selection bias. Moreover, mixed-format examinations enabled within-subjects comparisons of VSAQs and MCQs, under comparable instructional and assessment conditions. Standardizing scores further reduced variability and ensured reliable comparisons.

However, this study also has limitations. While GPA is an objective and widely used measure of academic achievement, it simplifies complex learning outcomes and may overlook critical thinking and skill development [19]. Additionally, the variability in content and distribution between VSAQs and MCQs within real-world examinations could introduce bias. However, we mitigated this by aggregating data from multiple examinations and cohorts, standardizing scores, and conducting analyses across two student populations.

Implications for practice and future research

Our findings indicate that VSAQs have a higher discriminative ability than MCQs, effectively distinguishing students both within individual examinations and across multiple examinations based on GPA. This suggests that VSAQs provide a more robust and valid question format for evaluating academic performance, supporting their implementation to enhance the discriminatory power and reliability of assessments. Incorporation of VSAQs into assessments can also enhance the ability to identify students in need of early interventions and support, while also recognizing those who excel. Moreover, VSAQs could play a valuable role in the selection process for medical studies by assessing study success potential while simultaneously providing an authentic preparation for the curriculum. Teachers could leverage VSAQs to implement tailored strategies for improving learning outcomes and to identify exceptional students for advanced opportunities. Future studies could expand academic performance measures beyond GPA, incorporating assessments of skill development and work-place learning. These additional metrics would provide a more holistic evaluation of the effectiveness of VSAQs and MCQs across different educational contexts. Additionally, exploring whether VSAQs are associated with long-term academic and professional success, including performance in real-world clinical settings, would provide deeper insights into their utility.

Conclusion

This study highlights the superior effectiveness of VSAQs over MCQs in distinguishing undergraduate medical students with varying academic performance levels. Integrating VSAQs into assessments enhances the discriminative power and robustness of assessments and may improve early identification of both poor and excellent performing students, allowing for targeted interventions, tailored support, and advanced opportunities. Future research could explore the broader applicability of VSAQs across diverse educational settings and assess their potential to predict long-term academic achievements.

References

1. Al-Rukban MO. Guidelines for the construction of multiple choice questions tests. Journal of Family & Community Medicine. 2006;13(3):125-33.
2. Schuwirth L, van der Vleuten C. Written Assessment. ABC of Learning and Teaching in Medicine: Wiley-Blackwell; 2017. p. 65-9.
3. 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.
4. Schuwirth LWT, Vleuten CPM, Donkers HHLM. A closer look at cueing effects in multiple-choice questions. Medical Education. 1996;30(1):44-9.
5. 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.
6. 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.
7. 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.
8. van Wijk EV, Janse RJ, Ruijter BN, Rohling JHT, van der Kraan J, Crobach S, et al. Use of very short answer questions compared to multiple choice questions in undergraduate medical students: An external validation study. PloS One. 2023;18(7):e0288558.
9. Sam AH, Field SM, Collares CF, van der Vleuten CPM, Wass VJ, Melville C, et al. Very-short-answer questions: reliability, discrimination and acceptability. Medical Education. 2018;52(4):447-55.
10. 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;Volume 10:943-8.
11. Mee J, Pandian R, Wolczynski J, Morales A, Paniagua M, Harik P, et al. An experimental comparison of multiple-choice and short-answer questions on a high-stakes test for medical students. Advances in Health Sciences Education. 2023;29(3).
12. Eijsvogels TMH, van den Brand TL, Hopman MTE. Multiple choice questions are superior to extended matching questions to identify medicine and biomedical sciences students who perform poorly. Perspectives on Medical Education. 2013;2(5):252-63.
13. Fenderson BA, Damjanov I, Robeson MR, Veloski JJ, Rubin E. The virtues of extended matching and uncued tests as alternatives to multiple choice questions. Human Pathology. 1997;28(5):526-32.
14. Frank J, Danoff D. The CanMEDS initiative: implementing an outcomes-based framework of physician competencies. Medical Teacher. 2007;29(7).
15. RemindoToets [Available from: https://www.paragin.nl/remindotoets/.
16. De Champlain AF. A primer on classical test theory and item response theory for assessments in medical education: Classical test theory and item response theory. Medical Education. 2010;44(1):109-17.
17. Krathwohl DR. A Revision of Bloom's Taxonomy: An Overview. Theory Into Practice. 2002;41(4).
18. Brookhart SM, Guskey TR, Bowers AJ, McMillan JH, Smith JK, Smith LF, et al. A Century of Grading Research. Review of Educational Research. 2016;86(4).
19. York TT, Gibson C, Rankin S. Defining and Measuring Academic Success. Practical Assessment, Research, and Evaluation. 2015;20(1).

Appendix

Appendix 1 – Supplemental Table 1

Supplemental Table 1. Distribution of VSAQs and MCQs across the themes within the examinations.

	RM 2021		RM 2022		RM 2023		DA 2022		DA 2023		DA 2024	
	VSAQ	MCQ	VSAQ	MCQ	VSAQ	MCQ	VSAQ	MCQ	VSAQ	MCQ	VSAQ	MCQ
Theme 1	2	1	1	1	1	3	2	4	4	3	4	1
Theme 2	5	7	6	9	5	5	4	5	4	7	3	3
Theme 3	23	6	13	11	14	13	2	4	2	5	4	1
Theme 4	4	11	5	8	5	10	3	4	6	3	7	1
Theme 5	5	6	3	6	4	6	3	8	2	7	4	2
Theme 6	3	1	1	1	2	1	2	11	6	4	5	4
Theme 7	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	5	7	3	5	1	4
Theme 8	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	1	5	3	6	3	4

RM=Regulation and Metabolism; DA=Diseases of the Abdomen; VSAQ=very short answer question; MCQ=multiple-choice question.

RM: Theme 1: Regulation of the temperature, Theme 2: Regulation of the reproduction, Theme 3: Regulation of the thyroid, Theme 4: Stomach, bowl and liver, Theme 5: Metabolism, Theme 6: Nutrition.

DA: Theme 1: Abdominal swelling, Theme 2: Stomach complaints, Theme 3: Jaundice, Theme 4: Acute abdominal pain, Theme 5: Chronic abdominal pain and defecation disorders, Theme 6: Blood loss, Theme 7: Anatomy, Theme 8: Other.

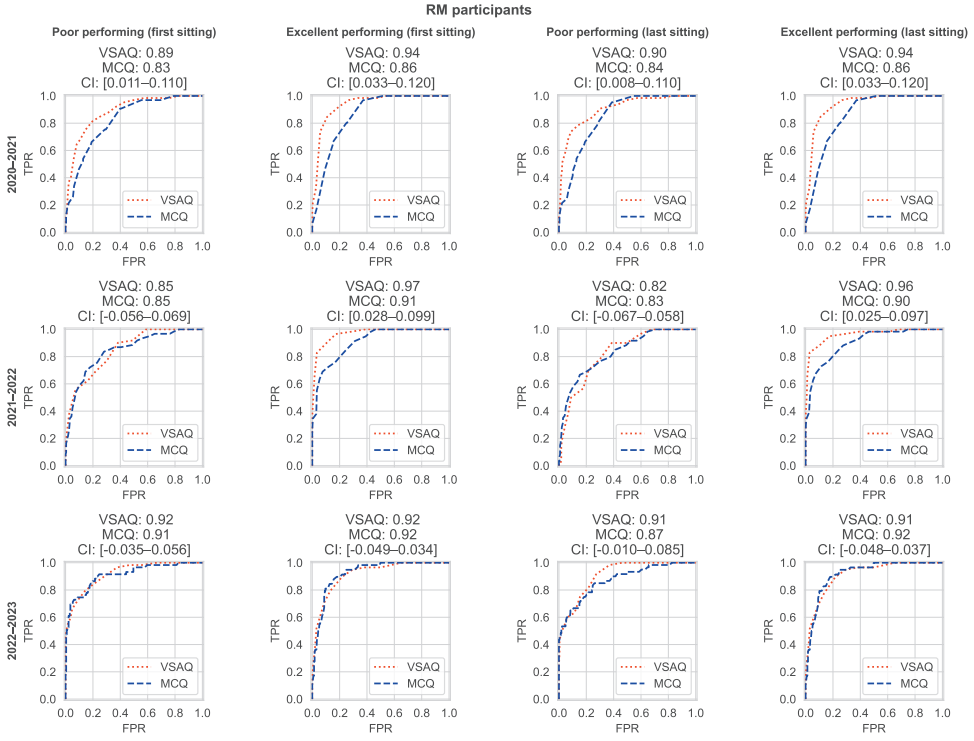
Appendix 2 – Supplemental Table 2

Supplemental Table 2. Distribution of Bloom's taxonomy levels of VSAQs and MCQs within examinations.

	RM 2021	RM 2022	RM 2023	DA 2022	DA 2023	DA 2024
Remember						
MCQ	43.8%	47.4%	52.5%	41.7%	42.5%	55.0%
VSAQ	80.9%	72.4%	77.4%	54.6%	40.0%	35.5%
Understand						
MCQ	53.3%	47.4%	40.0%	22.9%	27.5%	25.0%
VSAQ	14.3%	13.8%	12.9%	9.1%	13.3%	16.1%
Apply						
MCQ	3.1%	5.3%	7.5%	35.4%	30.0%	20.0%
VSAQ	4.8%	13.8%	9.7%	36.4%	46.7%	48.4%

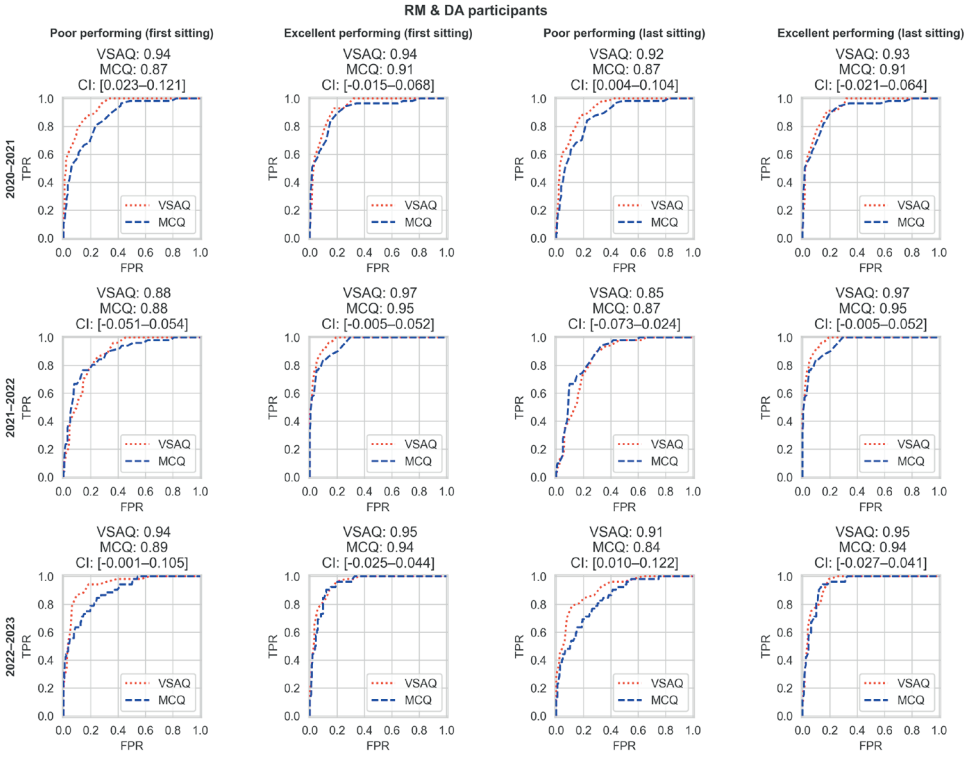
RM=Regulation and Metabolism; DA=Diseases of the Abdomen; VSAQ=very short answer question; MCQ=multiple-choice question.

Appendix 3 – Supplemental Figure 1

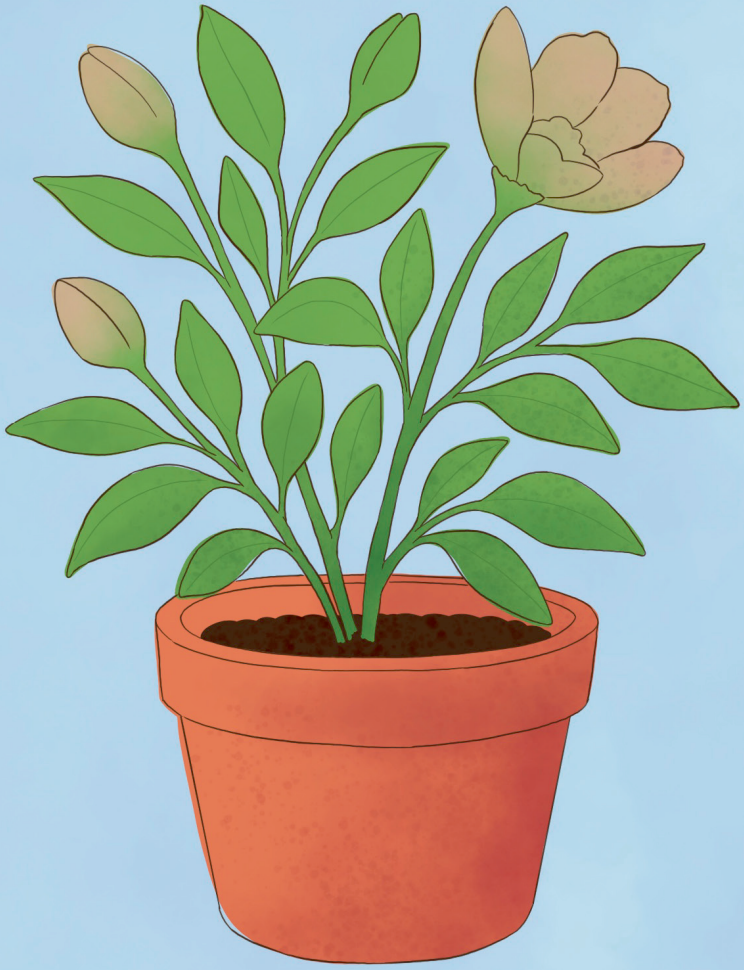


Supplemental Figure 1. Receiver Operating Characteristic (ROC) curves for both poor and excellent performing students using the first sitting or last sitting grades from the RM participants cohort 2020-2021 (RM 2021), cohort 2021-2022 (RM 2022), and cohort 2022-2023 (RM 2023). Red = VSAQs; Blue = MCQs; CI = 95% confidence interval; TPR = true positive rate; FPR = false positive rate. C-statistics are shown above each graph, which were calculated from the area under the ROC curve.

Appendix 4 – Supplemental Figure 2



Supplemental Figure 2. Receiver Operating Characteristic (ROC) curves for both poor and excellent performing students using the first sitting or last sitting grades from the RM & DA participants cohort 2020-2021 (RM 2021 & DA 2022), cohort 2021-2022 (RM 2022 & DA 2023), and cohort 2022-2023 (RM 2023 & DA 2024). Red = VSAQs; Blue = MCQs; CI = 95% confidence interval; TPR = true positive rate; FPR = false positive rate. C-statistics are shown above each graph, which were calculated from the area under the ROC curve.



Part I: Very short answer question

Chapter 4

The battle of question formats: A comparative study of retrieval practice using very short answer questions and multiple choice questions

Elise V. van Wijk
Mario de Jonge
Floris M. van Blankenstein
Roemer J. Janse
Alexandra M.J. Langers

BMC Medical Education. 2024;24(1):1547.
DOI: 10.1186/s12909-024-06538-0

Abstract

Background: Retrieval practice is a highly effective learning strategy that enhances long-term retention by encouraging the active recall of information. However, the optimal question format for maximizing knowledge retention remains uncertain. In this study, we compared the effect of very short answer (VSAQ) versus multiple-choice question (MCQ) practice tests on students' knowledge retention. By analyzing these two formats, we aim to identify the most effective approach to retrieval practice, thereby helping to optimize its implementation and improve learning outcomes.

Methods: In this randomized within-subjects study, students (n = 45) practiced with both VSAQs and MCQs in an extracurricular lifestyle course, without receiving feedback. The final retention test consisted of identical questions in both formats. A 2 × 2 repeated measures ANOVA was used to determine the effect of question format in practice testing and final test on final test score. Additionally, digital questionnaires were used to explore students' test-taking experiences.

Results: The VSAQs were answered incorrectly more frequently on the practice tests and final test. There was no main effect of practice question format on final test performance, and no interaction effect between question format on the practice and final test. Regardless of question format, most students thought the practice tests were beneficial for learning.

Conclusions: We found no evidence indicating that either MCQ or VSAQ is more effective for knowledge retention during retrieval practice. The lower initial retrieval success in the VSAQs, indicated by the higher degree of incorrect answers on the practice tests, might have limited their effectiveness during retrieval practice. To optimize the use of VSAQs in retrieval practice, it seems important to improve initial retrieval success to maximize learning outcomes.

Introduction

Retrieval practice is a highly effective learning strategy that enhances robust learning and long-term retention by requiring individuals to recall previously learned information from memory. This phenomenon, known as the testing effect or test-enhanced learning (TEL), is commonly implemented through practice tests or quizzes [1,2,3]. In health professions education, multiple-choice questions (MCQs) are frequently utilized as the practice test format. However, open-ended questions such as very short answer questions (VSAQs) are gaining popularity due to their advantageous psychometric properties and ability to provide deeper insights into students' misperceptions and common errors [4,5,6,7,8,9]. Despite the recognized learning benefits of retrieval practice, the question of which format – MCQs or open-ended questions – is most effective for enhancing learning and improving knowledge retention remains unresolved and is the subject of ongoing debate [10,11,12,13,14].

In health professions education, MCQs, mostly in the form of Single Best Answers (SBA), are commonly used to assess students due to their ease of marking, feasibility, and high reliability [15, 16]. However, their susceptibility to cueing, and more superficial recognition-based exam preparation [4, 7, 17] can decrease the effort of knowledge retrieval from memory, resulting in less effective retrieval practice and knowledge retention [18]. This aligns with the retrieval effort hypothesis, which states that retrieval requiring greater cognitive effort proves more beneficial for learning than easy retrieval with prompts or cues [19]. Consequently, students focusing on cues to respond to the question with the least effort (i.e., principle of least effort [20]), may fail to reproduce answers from memory on a later test without available answer options [21, 22].

Open-ended questions may be more effective for retrieval practice compared to MCQs, though the evidence is inconclusive [19]. Previous studies in cognitive psychology predominantly compared MCQs with open-ended Short Answer Questions (SAQs). Although SAQs allow for more extensive responses, such as several sentences, they are often employed in practice a way that mirrors VSAQs, with answers typically restricted to one or a few words. Several studies show that retrieval practice using SAQs is more effective than retrieval practice with MCQs, most likely because SAQs require more effort in retrieving information from memory and are typically experienced as more difficult [4, 6, 18, 23,24,25].

For example, Gay *et al.* [24] compared the effects of six SAQ and MCQ tests, covering identical concepts, in two groups of 14 students in an educational research course. The study assessed students' performance on a final exam, which tested the same concepts using both question formats. The results showed that SAQ practice led to better knowledge retention when tested with SAQs on the final exam, while the retention was comparable between the two formats when tested with MCQs on the final exam. This finding is in line with the theory of transfer-appropriate processing (TAP), which refers to the idea that learning and knowledge retention are improved if the type of processing used during retrieval matches the type of processing used during encoding, in this case the question format [21, 26]. Another study compared VSAQs and MCQs in a formative online test at the end of a pharmacology course and after one year [27]. They found an overall increase in knowledge retention after one year, which was higher in the students who started with VSAQs in the formative test. Nonetheless, this study mainly assessed the effect of the order in which the questions were given in the tests (first VSAQs or first MCQs) and which

question format offered a better preparation in basic pharmacology. There are also studies that found no advantages of retrieval practice using SAQs against MCQs, which may be associated with the lower level of initial retrieval success in SAQs [14, 28,29,30].

In summary, consensus on the impact of different question formats on knowledge retention remains elusive, and previous studies mainly tested in a non-medical or simulated setting [18, 23,24,25, 29]. Therefore, we aimed to investigate the effectiveness of retrieval practice with VSAQs and MCQs on knowledge retention in a real-life educational stetting among health science students. We also evaluated students' experiences with the practice tests. Our hypothesis is that VSAQs will be more effective compared to MCQs as practice format. In line with the TAP theory, we expect that the benefit of VSAQs over MCQs will be most pronounced when students are also tested with VSAQs on the final retention test. Additionally, this effect may be influenced by the greater sensitivity of VSAQs as a memory test, as they require active retrieval rather than recognition. By investigating this, we can further optimize the implementation of retrieval practice in health professions curricula and improve learning outcomes.

Methods

Setting

This study was conducted during the 2024 Students Experienced in Lifestyle and Food (SELF) course at Leiden University Medical Center (LUMC). The SELF-course is an extracurricular program that covers lifestyle medicine topics not covered in the standard medical curriculum. It is offered annually in February-March at all medical schools in the Netherlands to (bio)medical or healthcare related students, and early-career doctors. The course comprises eight evening sessions (i.e., masterclasses) of two hours every week, each focusing on a specific lifestyle-related theme (*Appendix 1 – Course overview*). This years' masterclasses were each divided into two lectures, given by a different speaker. The masterclasses were given live in a lecture hall, with an option for students to attend virtually. No preparatory work was required for the masterclasses, and the presentations of the speakers were only shared with the students at the end of the course. Directly following each masterclass, students completed a short formative quiz (i.e., practice test), with the two highest scoring students receiving a small reward from the SELF-board (e.g., a cookbook). The course concluded with a final assignment, where students created a poster to encourage healthier eating habits among general practitioner patients. Upon successful completion of the course, students were awarded a certificate. Medical students who participated in the honours program of the university received additional in-depth practical assignments related to the themes, but not directly to the lectures, and were awarded with study credits upon successful completion.

Participants

A total of 62 participants enrolled in the SELF-course, of 48 complete both the practice tests and the final test. We only included bachelor and master students enrolled in the course, and excluded participants who were already working in healthcare and no longer enrolled as students (n=2) to minimize potential confounding effects due to differences in baseline knowledge. Additionally, students who had missed more than two practice tests were excluded from the analysis (n=1), as they completed only a single practice test. This resulted in a final sample size of 45.

Study design and procedure

We conducted an experimental study with a within-subjects design, in a real-world educational setting. We provided information to the students during the first masterclass of the SELF-course, which they also received by e-mail (*Figure 1*). The practice test of this first masterclass was created by the SELF-board, and not part of our study to give students enough time to think about participation. Our research team developed the subsequent four practice tests. Following the second masterclass, participants were randomized by using a computer-generated sequence to start the practice test with either four VSAQs about the first lecture followed by four MCQs about the second lecture (*version A*), or the counterbalanced version (*version B*). Prior to the first practice test, students were asked for demographical data (*Appendix 2* includes both demographical questions and questions of the first practice test). In the subsequent masterclasses, the practice tests formats were reversed each week (Additional file 3 provides a figure with a more detailed overview of the different versions). To isolate the effects of retrieval practice of the question format, students were not provided with the correct answers or the presentations after the tests. The final test was administered at the beginning of the last masterclass in week 8. The retention intervals for the different practice tests were six weeks for the first practice test, five weeks for the second, four weeks for the third, and three weeks for the last practice test (*Figure 1*), and consisted of questions used in the four practice tests (n=32; 16 VSAQs, 16 MCQs). Within the two groups (randomized to version A and B) students were further randomized to take one of the two counterbalanced versions of the final test (*version 1 or 2*). In version 1, the first two questions of each of the masterclass lectures were presented in VSAQ format, while the last two questions were in MCQ format (VSAQ | MCQ, *Figure 1 & Appendix 3 – Supplemental Figure 1*).). In version 2, this order was reversed, with the first two questions were in MCQ format and the last two in VSAQ format (MCQ | VSAQ, *Figure 1 & Appendix 3 – Supplemental Figure 1*). Since students received half of the final test questions in the same format as those in the practice test (congruent questions) and the other half in a different format (incongruent questions), we were able to assess the theory of TAP. After completing the final test, students competed a digital questionnaire regarding their experiences, and they received the correct answers of the final test.

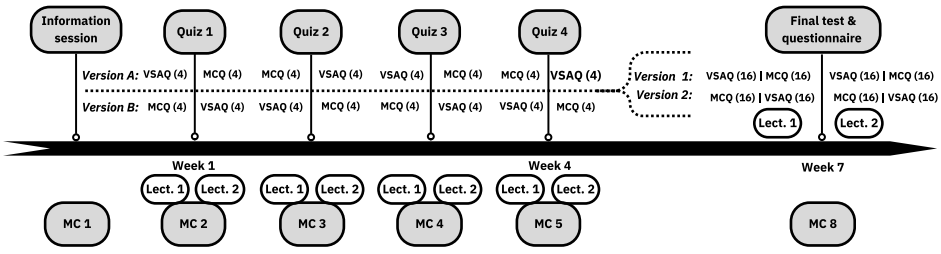


Figure 1. Study design with the different versions of the practice tests and final test. MC = masterclass; Quiz = practice test; Lect. = lecture; MCQ = multiple-choice questions; VSAQ = very short answer questions. The number of questions on the tests is reported between brackets.

Development of practice tests and final test

The principal investigator (EvW) developed the practice test questions based on lecture materials, additional information and key messages provided by the speakers. Two educationalists experienced in assessment question writing along with the speakers, reviewed the questions to ensure their quality and alignment with the lecture content. The educationalists discussed their suggestions with the principal investigator, which led to adjustments of the original questions. All questions were designed to assess the first two stages of the concept of Bloom's Taxonomy: recall and understanding [31]. Each question was designed to suit both the VSAQ and MCQ format, without altering the question's structure or content. Students received the questions either in VSAQ or MCQ format. The MCQs had four answer options, with one correct answer (i.e., single best answer questions). Answer options were randomly ordered for each participant. The VSAQs required responses of no more than four words.

Scoring procedure

MCQs were automatically scored by using a standard letter key. For the VSAQs, a predefined list of acceptable answers was used, and these were automatically marked as correct in the digital assessment system. This predefined list was created by the item writer (EvW) together with the speakers. Subsequently, all incorrect answers were reviewed, and correct answers not initially included in the predefined list were added. Each correct answer was awarded one point, while incorrect or partially correct questions yielded no point, with a maximum score of 8 point on each practice test and 32 on the final test. Three investigators of the research team with a (bio)medical and educational background (EvW, MdJ, and AL) scored the VSAQs independently in a blinded manner. They indicated which of the answers for each question in the list they thought was correct, incorrect, or which of the answers they were not sure about. The individual scores were then reviewed and discussed a joint session to reach consensus on the final scores. Out of the total 2683 answers, only 77 responses (2.87%) required further discussion to reach consensus. If the question explicitly required a single response but multiple answers were given, the response was marked incorrect, even if one of the answers was correct. Conversely, for questions where the number of expected answers was not specified, the response was considered correct if at least one of the provided answers was correct. We also approved answers that were not explicitly mentioned during the lectures, and were thus not included in the predefined list of answers, but proved to be correct after all. One question, in both question formats, was removed during the reviewing process because the topic was not discussed during the lecture. The speakers were consulted about the answers that required more in-depth knowledge of the topic.

Data collection

Directly at the end of the second, third, fourth, and fifth masterclass, participants undertook digital practice tests via the Remindo assessment system [32]. These tests could be taken either in the lecture hall or remotely, as approximately 20% of students attended the masterclasses online. For two students who were unable to access the digital system, paper-based tests were provided. Each practice test consisted of eight questions, from which the first four questions focused on the content from the first lecture, followed by four questions related to the second lecture. The order of the questions was fixed, ensuring a structured progression through the material covered in each lecture. On the digital practice tests, students were required to answer all questions, without the option to skip questions or to return to previous questions.

All students who participated in the practice tests completed every question on both the digital and paper version. This setup was consisted across all practice tests administered during the study. Participants were instructed not to use study materials during the test, but we did not use proctoring to verify this. We allocated 20 min to finish each practice test, after which all students were allowed to leave. The final test, administered at the start of the eighth masterclass, consisted of all questions derived from the four practice tests (*n*=32; 16 VSAQs and 16 MCQs). Upon initiation and completion of the final test, students answered questions in the digital assessment system regarding their experiences with the practice tests. For this study we analysed the four questions answered before the final test. The complete questionnaire can be found in *Appendix 4*.

Data analysis

Descriptive statistics for the demographics and previous experience of participants were calculated. Continuous data, depending on their distribution, are reported as mean and standard deviation (SD) or median and interquartile range (IRQ). Categorical variables are expressed as a number (percentage). Total test scores were calculated by adding the points for each question. In line with previous research, we did not correct for guessing on the MCQs [14, 24, 27]. We reported the percentages of questions practiced in VSAQ or MCQ format that were incorrect/correct in the practice test and/or final test, and used a 2 (practice test format) × 2 (final test format) repeated measures ANOVA to analyse the main effects and interaction effect of question formats on practice and final test on final test score. The answers to the short questionnaire regarding student's experiences were reported as answer distributions. Missing data were not imputed. Statistical significance is denoted by p-values and 95% confidence intervals, with a p-value of less than 0.05 considered significant. All statistical analyses were performed using R version 4.3.1 (R Foundation for Statistical Computing, Vienna, Austria).

Ethical approval

This study was approved by the LUMC Educational Research Review Board (OEC/ERRB/20231010/1). Participation in this study was voluntary and all students received verbal and written information prior to the study. Upon initiation of the first practice test, students provided informed consent.

Results

Demographics

Of the 62 participants in the SELF-course, 55 participants completed the practice tests. The final retention test was completed by 48 participants. We excluded three participants, because they were not a student (*n*=2), or had only completed less than two practice tests (*n*=1), resulting in a sample of 45 students (*Figure 2*). *Table 1* shows that the majority of participants studied medicine (71.1%), and were bachelor students (57.7%). Two students prepared for the final test. There were no differences in characteristics between the student groups in the counterbalanced versions (*version A and B*).

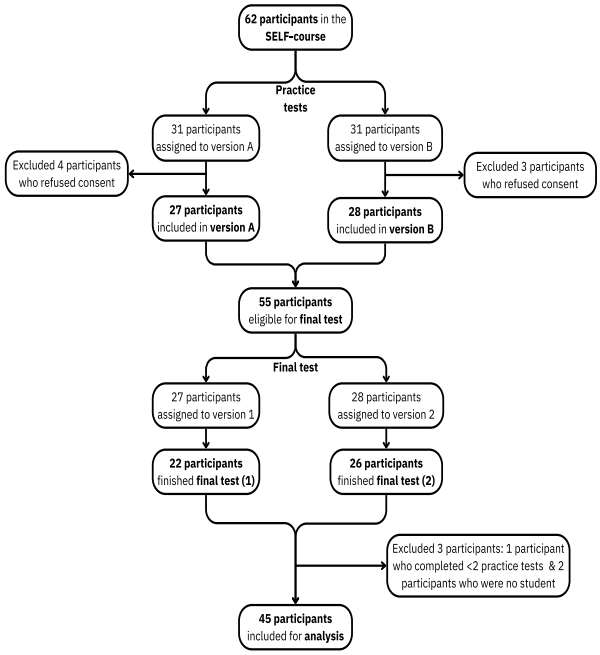


Figure 2. Flowchart of participants of the study

Table 1. Characteristics of the participants of the study

	Overall (*n*=45)
Age, median (IQR)	21 (20, 23)
Study, n (%)	
Biomedical Sciences	7 (15.6)
Medicine	32 (71.1)
Psychology	2 (4.4)
Vitality and Aging	3 (6.7)
Biomedical Sciences & Medicine	1 (2.2)
Study year, n (%)	
Bachelor[a] year 1	4 (8.9)
Bachelor year 2	11 (24.4)
Bachelor year 3	11 (24.4)
Master[b] year 1	9 (20.0)
Master year 2	4 (8.9)
Master year 3	3 (6.7)
Waiting period[c]	3 (6.7)
Honours program	
Yes	15 (33.3)
Previous lifestyle courses	
Yes	4 (8.9)

IQR = interquartile range

[a]Bachelor = pre-clinical phase in medicine

[b]Master = clinical phase in medicine

[c]Waiting period before the clinical clerkships during the postgraduate phase.

Performance on practice tests and final test

The mean total score on the final test was 14% lower than the mean overall score of all four practice tests combined (mean percentage score (SD) practice tests: 66.2% (12.1) vs. final test: 52.7% (10.1); paired t-test: $t(44) = 6.99$; $p < 0.01$; 95%CI [0.10,0.17]). Successful retrieval practice, defined as correctly answered questions in both the practice test and the final retention test, occurred in 50.4% of questions practiced in the MCQ format vs. 39.8% of questions practiced in VSAQ format (*Table 2*). For both question formats approximately half of the initially correct answers were forgotten on the final test (23.5/50.4% for MCQs; 18.8/39.8% for VSAQs). 23.5% of the MCQs, and 18.8% of VSAQs were initially answered correctly on the practice test, but answered incorrectly in the final test. MCQs were more often answered correctly on the practice tests (73.9% vs. 58.6%), which is not surprising and was to be expected since MCQs also allow for guessing.

Table 2. Percentages of correct and incorrect answers on VSAQs and MCQs

Practice test question format and performance	Final test performance[a]				
	Correct		Incorrect		Total
	MCQ	VSAQ	MCQ	VSAQ	
Multiple-choice questions					
Correct	28.3%	22.1%	8.7%	14.8%	73.9%
Incorrect	4.0%	2.7%	8.9%	10.5%	26.1%
Very short answer questions					
Correct	21.1%	18.7%	7.7%	11.2%	58.6%
Incorrect	7.2%	4.2%	13.7%	16.3%	41.4%

[a]Questions identical to practice test questions with a random 50/50 distribution of both VSAQs and MCQs

Effect of question format on final retention test performance

The 2 × 2 (question format practice test × question format final test) repeated measures ANOVA showed no significant main effect of question format in practice tests on the final test score ($F(1,44) = 3.23$, $p = 0.08$, $\eta 2 = 0.02$). The mean proportion of correctly answered questions on the final test was similar for the content practiced with MCQs (M (SD) 57.1% (2.2)) and VSAQs (51.2% (2.3)). Question format on the final test had a significant main effect on the final test score ($F(1, 44) = 22.80$, $p < 0.01$, $\eta 2 = 0.10$) with lower proportional scores on VSAQs compared to MCQs, (M (SD) MCQ: 61.0% (2.2), VSAQ: 47.2% (2.1)). Figure 3 shows the proportion of correctly answered questions on the final test as a function of question format in the practice tests and on the final test (*Appendix 5* shows the table with the mean percentages). There was no interaction effect between question format on practice test and final test ($F(1, 44) = 0.16$, $p = 0.69$, $\eta 2 < 0.01$).

Student's experiences

Of all students, 68.9% (n=31) students thought the practice tests helped them remember the content of the lectures better. There was no difference between the estimation of how much students thought they still remembered of the lectures practiced with VSAQs vs. MCQs (median of 0%, *Appendix 6* shows the histogram).

Discussion

In this study we investigated the effectiveness of retrieval practice using MCQs and VSAQs on knowledge retention in an educational setting. We found no significant difference in knowledge retention between retrieval practice with MCQs and VSAQs, and no interaction effect between the question formats. This indicates that we found no evidence of an advantage associated with either question format in retrieval practice. However, VSAQs were more challenging, as reflected by lower scores on both practice and final tests. Most students reported that the practice tests were beneficial, regardless of the question format.

Our findings align with previous research comparing SAQs and MCQs, which also found no differences in the effectiveness of retrieval practice [14, 28,29,30, 33]. Similarly, we observed no significant interaction effect between practice and final test question format, indicating that matching question formats does not enhance the retrieval effect. While it challenges the TAP theory, it supports earlier findings suggesting that retrieval practice benefits do not depend on an exact match between conditions in practice and final test [11, 12, 14, 30, 33].

One explanation for the lack of a retrieval benefit in our study could relate to the level of initial retrieval success, which is often lower for VSAQs due to their greater difficulty [3, 14, 23, 34]. Rowland's meta-analysis [34] indicates that higher retrieval success often leads to a stronger retrieval effect. If initial retrieval is unsuccessful, it may lead to lower levels of knowledge retention [14]. However, a recent study suggests there may be no consistent relationship between initial retrieval success and the testing effect [35], underscoring the complexity of this relationship and the potential influence of other moderating factors, such as individual differences and contextual variations. Feedback might correct for unsuccessful retrieval, but we intentionally did not provide feedback after the practice tests to isolate the direct effects of the different question formats. Consistent with earlier findings [4, 6, 7, 27], VSAQs were answered incorrectly more often than the MCQs, resulting in lower retrieval success.

Previous studies show mixed results when corrective feedback is provided, with some indicating better learning outcomes for SAQs compared to MCQs, while others found no differences [11, 12, 30, 33]. Although feedback ensures exposure to correct answers, it also introduces indirect effects that are often not examined, complicating the assessment of direct retrieval effects of different question formats on learning [3, 36]. The ability to recognize the correct answer among MCQ options may also serve as implicit feedback, potentially enhancing the learning effect for MCQs. In studies showing greater knowledge retention with (V)SAQs compared to MCQs, practice and final test questions did not exactly match, and it was unclear whether feedback was provided or what the initial retrieval success levels were [24, 27]. The timing of the practice tests may further influence retrieval success. While administering them shortly after learning, as in our study, may enhance success, scheduling them late in the evening after a two-hour lecture might have hindered it, particularly for the more challenging VSAQs. In sum, and in line with previous research [3, 30, 33, 35, 37], finding the appropriate balance between retrieval success and effort is complex, as it may also be influenced by various other factors.

Strengths & limitations

To our knowledge, this study is the first to compare the effect of VSAQ and MCQ practice tests among health science students using a controlled design within a real-life educational setting.

We used a rigorous methodology, including a within-subjects design, counterbalanced test versions, and identical questions in both tests. Although the use of identical questions across tests is not common practice in a real-life educational setting, it was necessary to ensure a pure measurement of the memorial effects of question formats. Students had comparable prior knowledge due to the absence of prior mandatory education on lifestyle topics, and most had not taken previous lifestyle courses. Moreover, there were no external incentives, such as preparatory assignments, that could influence learning effort. This design allowed us to directly measure retrieval practice effects without simulating a test environment [11, 18, 23]. The random distribution of both question formats in the final retention test equalized difficulty levels, and enabled analysis of how differences between initial and final test formats affected performance.

The absence of feedback, while allowing a direct measurement of retrieval effects, may have influenced our results, particularly for VSAQs, were students showed lower initial retrieval success. Research indicates that feedback can enhance the testing effect by reinforcing correct responses and correcting errors, especially when initial retrieval success is unsuccessful [11, 12, 34, 36, 38]. Including feedback might have amplified the differences between question formats, improving VSAQ performance more significantly than MCQ performance. Additionally, the limited number of test questions and sample size, though larger than in most prior studies [23, 24, 29], may have affected the robustness of our findings. A larger sample size could reveal differences between MCQ and VSAQ formats, warranting further investigation.

While our within-subjects design and random assignment of students to counterbalanced versions minimized confounding variables, such influences cannot be entirely eliminated in a real-world educational setting. Nonetheless, our study design more accurately reflects the uncontrolled conditions typical of real-world educational settings, enhancing the generalizability of the findings. Variations in lecture attendance (online vs. in-person) and test format (digital vs. paper) might have influenced student motivation and test performance. Furthermore, the voluntary, low-stakes nature of this course may differ from traditional graded courses, potentially affecting both students' engagement with the material and their test performance. However, this low-stakes setting reduced additional studying between tests, providing a clearer view of retrieval practice effects. It remains possible, though, that some motivated students independently sought feedback. Because we compared the effect of two different question formats on retrieval practice without including a control group of students who did not complete practice questions, we cannot conclude that learning was directly supported by the practice questions in our setting. However, the general effectiveness of practice testing has been consistently demonstrated in previous research.

Implications for practice and future research

Our findings suggest no clear advantage of either question format in enhancing learning through practice tests, regardless of the format used in the final test. Achieving an optimal balance between retrieval effort and retrieval success may enhance the effectiveness of retrieval practice. Practically, this indicates that teachers can choose either VSAQs or MCQs for practice testing, as performance was not influenced by practice question format or alignment with the final test format. However, VSAQs may offer learners better insights into knowledge gaps and misperceptions compared to MCQs, making them a preferred diagnostic

tool to guide future learning [8, 9], particularly when assessing higher-order skills such as knowledge application. In educational settings, VSAQs could be prioritized when the goal is to address specific misconceptions, as they provide deeper insights into student understanding. For lower-order skills like remembering and understanding, as examined in this study, VSAQs are more likely to reflect retrieval failures or memory interference, which can be valuable for formative assessments but may require additional feedback to maximize their effectiveness.

A key advantage of VSAQs is their ability to reduce 'foresight bias' common with MCQs, providing a more reliable estimation of students' understanding, and better predicting retrieval success [39, 40]. This can help teachers identify underperforming students earlier and intervene more effectively. Additionally, VSAQs eliminate the need to create plausible distractors, a challenging and time-consuming aspect of MCQ construction [41,42,43,44], thereby saving teachers time and improving test quality. To address teachers' limited experience with VSAQs, providing clear instructions or workshops on constructing effective VSAQs would be beneficial.

To enhance the effectiveness of retrieval practice, particularly with VSAQs, we recommend integrating immediate or delayed feedback [11, 12, 45], and implementing repeated spaced practice retrieval [37]. As alternative for teacher-learner feedback, self-assessment by students in immediate self-feedback VSAQs might be a good option to use in formative assessments [9]. In this format, students give self-feedback on the understanding of the correct answers, which can help students recognize their knowledge gaps and guide further learning. Hybrid questions, such as combining VSAQs with MCQs or using stepwise MCQs, offer another promising approach [38, 39, 46]. These formats engage students in initial effortful retrieval (i.e., answering an open question without cues) followed by multiple-choice options for direct feedback, which balances retrieval effort and success. Further research is needed to evaluate the effectiveness of these formats in health professions education. Moreover, future research could explore factors influencing the effectiveness of retrieval practice with different question formats, including the level of initial retrieval success, individual differences in memory strength, learning strategies, and student motivation. Studies in diverse educational contexts, with longer retention intervals, or questions that assess higher levels of cognitive learning could further unravel the impact of retrieval practice on learning outcomes across different question formats.

Conclusion

We found no evidence to suggest that either question format, MCQ or VSAQ, is more effective for knowledge retention through retrieval practice. Despite higher retrieval effort in VSAQs, their lower initial retrieval success may have limited their effectiveness in enhancing retention. Nevertheless, practice testing with VSAQs offers valuable insights into knowledge gaps and provides a more reliable estimation of students' understanding. To optimize learning outcomes and enhance knowledge retention, it seems important to increase initial retrieval success, which could be achieved through feedback or repeated practice sessions.

References

1. Dunlosky J, KA R, Marsh E, Nathan M, Willingham D. Improving Students' Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology - PubMed. Psychological science in the public interest : a journal of the American Psychological Society. 2013 Jan;14(1).
2. 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.
3. Karpicke JD. Retrieval-based learning: A decade of progress.: Academic Press; 2017.
4. Sam AH, Field SM, Collares CF, van der Vleuten CPM, Wass VJ, Melville C, et al. Very-short-answer questions: reliability, discrimination and acceptability. Medical Education. 2018;52(4):447-55.
5. 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.
6. van Wijk EV, Janse RJ, Ruijter BN, Rohling JHT, van der Kraan J, Crobach S, et al. Use of very short answer questions compared to multiple choice questions in undergraduate medical students: An external validation study. PloS One. 2023;18(7):e0288558.
7. 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.
8. 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.
9. 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).
10. Greving S, Richter T. Examining the Testing Effect in University Teaching: Retrievability and Question Format Matter. Frontiers in Psychology. 2018;9.
11. Kang SHK, McDermott KB, Roediger HL. Test format and corrective feedback modify the effect of testing on long-term retention. European Journal of Cognitive Psychology. 2007;19(4-5):528-58.
12. Little JL, Bjork EL, Bjork RA, Angello G. Multiple-Choice Tests Exonerated, at Least of Some Charges: Fostering Test-Induced Learning and Avoiding Test-Induced Forgetting. Psychological Science. 2012;23(11):1337-44.
13. Mcdaniel MA, Roediger HL, Mcdermott KB. Generalizing test-enhanced learning from the laboratory to the classroom. Psychonomic Bulletin & Review. 2007;14(2):200-6.
14. Smith MA, Karpicke JD. Retrieval practice with short-answer, multiple-choice, and hybrid tests. Memory. 2014;22(7):784-802.
15. Al-Rukban MO. Guidelines for the construction of multiple choice questions tests. Journal of Family & Community Medicine. 2006;13(3):125-33.
16. Schuwirth L, van der Vleuten C. Written Assessment. ABC of Learning and Teaching in Medicine: Wiley-Blackwell; 2017. p. 65-9.
17. Schuwirth LWT, Vleuten CPM, Donkers HHLM. A closer look at cueing effects in multiple-choice questions. Medical Education. 1996;30(1):44-9.
18. Larsen DP, Butler AC, Roediger III HL. Test-enhanced learning in medical education. Medical Education. 2008;42(10):959-66.
19. 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.
20. Samuels SJ. Effects of pictures on learning to read, comprehension and attitudes. Review of Educational Research. 1970;40(3):397-407.
21. Morris CD, Bransford JD, Franks JJ. Levels of Processing versus Tranfer Appropriate Processing. Journal of Verbal Learning and Verbal Behavior. 1977;16(5):519-33.
22. Roediger HL, Gallo DA, Geraci L. Processing approaches to cognition: The impetus from the levels-of-processing framework. Memory. 2002;10(5-6):319-32.
23. 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.
24. Gay LR. The comparative effects of multiple-choice versus short-answer tests on retention. Journal of Educational Measurement. 1980;17:45-50.
25. Greving S, Richter T. Practicing retrieval in university teaching: short-answer questions are beneficial, as multiple-choice questions are not. Journal of Cognitive Psychology. 2022;34(5):657-74.
26. Veltre MT, Cho KW, Neely JH. Transfer-appropriate processing in the testing effect. Memory. 2015-11-17;23(8).
27. Neumann J, Simmrodt S, Teichert H, Gergs U. Comparison of Online Tests of Very Short Answer versus Single Best Answers for Medical Students in a Pharmacology Course over One Year. Education Research International. 2021;2021:1-10.
28. Moreira BFT, Pinto TSS, Starling DSV, Jaeger A. Retrieval Practice in Classroom Settings: A Review of Applied Research. Frontiers in Education. 2019;4.
29. 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.

30. Bloom BS. Taxonomy of Educational Objectives: The Classification of Educational Goals: Longmans, Green; 1956 1956. 240 p.
31. RemindoToets [Available from: https://www.paragin.nl/remindotoets/.
32. McDaniel MA, Wildman KM, Anderson JL. Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. Journal of Applied Research in Memory and Cognition. 2012;1(1).
33. McDermott KB, Agarwal PK, D'Antonio L, Roediger HL, McDaniel MA. Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. Journal of Experimental Psychology: Applied. 2014;20(1).
34. Rowland CA. The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. Psychological Bulletin. 2014;140(6).
35. Green ML, Moeller JJ, Spak JM. Test-enhanced learning in health professions education: A systematic review: BEME Guide No. 48. Medical Teacher. 2018-4-3.
36. Vaughn KE, Rawson KA, Pyc MA. Repeated retrieval practice and item difficulty: Does criterion learning eliminate item difficulty effects? Psychonomic Bulletin & Review. 2013;20(6).
37. van den Broek GSE, van Gog T, Jansen E, Pleijsant M, Kester L. Multimedia Effects During Retrieval Practice: Images That Reveal the Answer Reduce Vocabulary Learning. Journal of Educational Psychology. 2021;113(8):1587-608.
38. Koriat A, Bjork RA. Illusions of Competence in Monitoring One's Knowledge During Study. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2005;31(2):187-94.
39. van den Broek GSE, Gerritsen SL, Oomen ITJ, Velthoven E, van Boxtel FHJ, Kester L, et al. Optimizing Multiple-Choice Questions for Retrieval Practice: Delayed Display of Answer Alternatives Enhances Vocabulary Learning. Journal of Educational Psychology. 2023;115(8):1087-109.
40. Little JL, Frickey EA, Fung AK. The role of retrieval in answering multiple-choice questions. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2019;45(8).
41. Little JL, Bjork EL. Optimizing multiple-choice tests as tools for learning. Memory & Cognition. 2015;43(1):14-26.
42. Gierl MJ, Bulut O, Guo Q, Zhang X. Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review. Review of Educational Research. 2017;87(6).
43. Ryan AT, Judd T, Wilson C, Larsen DP, Elliott S, Kulasegaram K, et al. Timing's not everything: Immediate and delayed feedback are equally beneficial for performance in formative multiple-choice testing. Medical Education. 2024/07/01;58(7).
44. Park J. Learning in a New Computerized Testing System. Journal of Educational Psychology. 2005;97(3).
45. Park J, Choi B-C. Higher retention after a new take-home computerised test. British Journal of Educational Technology. 2008;39(3).

Appendix

Appendix 1 – Course overview

Theme 1: Sleep

Lecture 1

- What is good sleep?
- What is the influence of sleep on (development of) health/disease?
- What is the influence of too little or too much sleep on the development of disease?
- What feasible advice can you give to patients regarding sleep?

Lecture 2

- What is the influence of caffeine on sleep and our biological clock?
- What is the influence of night shifts on our health?
- What is the influence of the other pillars (movement, nutrition, relaxation, mental health) on sleep?
- To what extent does our biological clock influence health and disease? And what is the role of the type of light (natural and artificial) on this?
- What is the best approach when you work night shifts? Is this universal or does it vary per person?

Theme 2: Mental Health

Lecture 1

- How do psychiatric medication and lifestyle interact?
- What is the influence of exercise on psychiatric disorders?
- What is the relationship between nutrition and mood?
- How can you use this information as a doctor in the clinic?

Lecture 2

- What is SOLK/ALK? (SOLK stands for "Somatisch Onvoldoende verklaarde Lichamelijke Klachten" in Dutch, which translates to "Somatically Unexplained Physical Complaints". ALK stands for "Aanpassingsstoornis met Lichamelijke Klachten", which translates to "Adjustment Disorder with Physical Complaints".)
- What is the relationship between emotion and SOLK/ALK complaints?
- How do you deal with patients with SOLK/ALK and what are the treatment options and advice?
- How can you as a doctor make use of this?
- What are the experiences of patients with these complaints?

Theme 3: Nutrition

Lecture 1

- What role does (highly) processed food play in disease?
- What concrete nutritional advice should you give in clinical practice?
- How do you deal with varying information on the internet about nutritional advice?
- What is your opinion about a healthy diet?

Lecture 2

- What is the connection between plant-based food and lifestyle and how can this be optimized for a healthy lifestyle?
- How can one transition to a plant-based diet?
- What are the ethical and medical-legal considerations regarding (plant-based) nutrition?

Theme 4: Microbiome and food transition

Lecture 1

- What is the microbiome and what is its function?
- What is the relationship between the microbiome, health, and disease?
- What are the effects of diet and lifestyle on the microbiome and are there lifestyle interventions that can improve the microbiome?
- What are recent scientific developments in the field of the microbiome?

Lecture 2

- No learning goals, because of a last-minute schedule change

Theme 5: Drugs and addiction

Lecture 1

- The student understands what addiction means both in practice as the working mechanism in the brain.
- The student understands how an addiction affects daily life.
- The student has insight into how sex and gender identity influence addiction.
- The student has insight into how to be careful with prescribing addictive substances between different sexes and gender identities.
- The student is able to discuss the topic of addiction with people of different sexes and gender identities.

Lecture 2

- The student understands the determinants of why people start vaping.
- The student recognizes the risk group that starts vaping.
- The student knows the dangers and physical harm to the body caused by vaping.
- The student is able to discuss the topic of vaping in the consultation room, also with the youth.
- The student is aware of the possible treatments to stop vaping.
- The student understands the urgency to stop the use of nicotine and vapes in the population.

Theme 6: Young and old

Lecture 1

- What are the first 1000 days?
- Which nutrition, lifestyle, and parenting factors influence development during the first 1000 days?
- What advice should be given to future parents in the consultation room, among others in the field of nutrition?

Lecture 2

- What is the effect of lifestyle on aging?
- How can lifestyle interventions ensure healthy aging?
- What are blue zones?
- What is the influence of the lifestyle of people living in blue zones on aging?

Theme 7: Movement

Lecture 1

- What is the effect of movement on health?
- What role does movement play in the development of diseases and disease prevention?
- What kind of lifestyle interventions are there in the field of movement, and how can these be implemented?

Lecture 2

- What is the importance of movement in aging?
- How does movement lead to better health outcomes and quality of life in the elderly?
- How can peer coaching be used for movement interventions and healthier aging?
- What do you advise patients/elderly in the consultation room about movement?
- What are the movement norms? Do people in the Netherlands move enough?
- What do you advise patients in the consultation room about movement?

Theme 8: In practice

Lecture 1

- How does behavioural change work?
- Why is behavior ('bad' habits) so difficult to change?
- Why do behavioural changes come to one patient and not to another?
- What makes lifestyle interventions effective or ineffective?
- What is the best approach for behavioural change for the individual? And from the government?

Lecture 2

- Bringing together all discussed topics and lifestyle pillars and concretizing with advice for practice.
- Do's and don'ts in the consultation room in the field of lifestyle.
- How can you as a doctor best influence the lifestyle behavior of patients?
- How do you deal with lifestyle problems in the consultation room?
- Cases from practice: good and bad examples.

Appendix 2 – Practice test and demographical questions

Practice test 1: Mental Health

Welcome to the first quiz of the SELF-course!

The quiz consists of eight questions in total, of which four are multiple-choice questions and four are very short answer questions (VSAQs). For the VSAQs, you are expected to give a short answer (maximum of four words). This first test will be preceded by a number of questions about personal data. We would like to use this for the analyses of the research. If you agree with the informed consent (indicate below), your (anonymized) data will be used for our study. If you do not agree, your data will only be used for the course and scoring.

It is not allowed to use study materials while taking the test. Good luck!

Demographical questions

1. What is your age?
2. What is your gender?
    - o Female
    - o Male
    - o Non-binary

3. What do you study?
    - o Biomedical Sciences
    - o Medicine
    - o Clinical Technology
    - o Psychology
    - o Nursing
    - o Vitality and Aging
    - o I already finished my study, and work as a:
    - o Other, namely:

4. Which study year are you in?
    - o Bachelor year 1
    - o Bachelor year 2
    - o Bachelor year 3
    - o Master year 1
    - o Master year 2
    - o Master year 3
    - o Waiting period before clinical clerkships
    - o Not applicable

5. Do you participate in the honours program?
    - o Yes

- o No

6. Did you participate in previous SELF-courses or other courses related to prevention and life style?
   - o Yes, how many SELF- or orther lifestyle courses did you follow?
   - o No

Version A

Very short answer questions

You have to answer these questions with a very short answer (one to four words)

1. Which lifestyle factor has both a preventive protective effect on anxiety disorders and an efficacy in the treatment of anxiety disorders?
2. Which lifestyle intervention is mentioned as the first treatment step in the guideline for depressive disorders?
3. Which category of psychopharmaceuticals has the highest chance of obesity as a side effect?
4. What type of diet is associated with an increased risk of mental health issues?

Multiple-choice questions

1. In about 40% of the complaints, no clear physical cause is found in the general practice. In what percentage of cases do such complaints become chronic?
   - A. 10%
   - B. 30%
   - C. 50%
   - D. 70%

2. Pain is an important protective mechanism of the brain. What is it trying to protect you from?
   - A. Emotional stress
   - B. Physical damage
   - C. Physical fatigue
   - D. Expected danger

3. What is the first intervention for patients suspected of having SOLK/ALK?
   - A. Additional examination
   - B. Psycho-education
   - C. Lifestyle interventions
   - D. Referral to a specialist

4. There is a theory that suggests that the brain actively makes estimates about upcoming sensory input instead of passively registering it. This could play a role in the development of chronic pain complaints. What is this theory called?
   - A. Extended cognition
   - B. Neurophenomenology

C. Predictive coding
D. Unconscious inference

Version B

Multiple-choice questions

1. Which lifestyle factor has both a preventive protective effect on anxiety disorders and an efficacy in the treatment of anxiety disorders?
   A. Activity
   B. Sleep
   C. Quit smoking
   D. Diet

2. Which lifestyle intervention is mentioned as the first treatment step in the guideline for depressive disorders?
   A. Activity
   B. Sleep
   C. Quit smoking
   D. Diet

3. Which category of psychopharmaceuticals has the highest chance of obesity as a side effect?
   A. Antidepressants
   B. Antipsychotics
   C. Benzodiazepines
   D. Stimulants

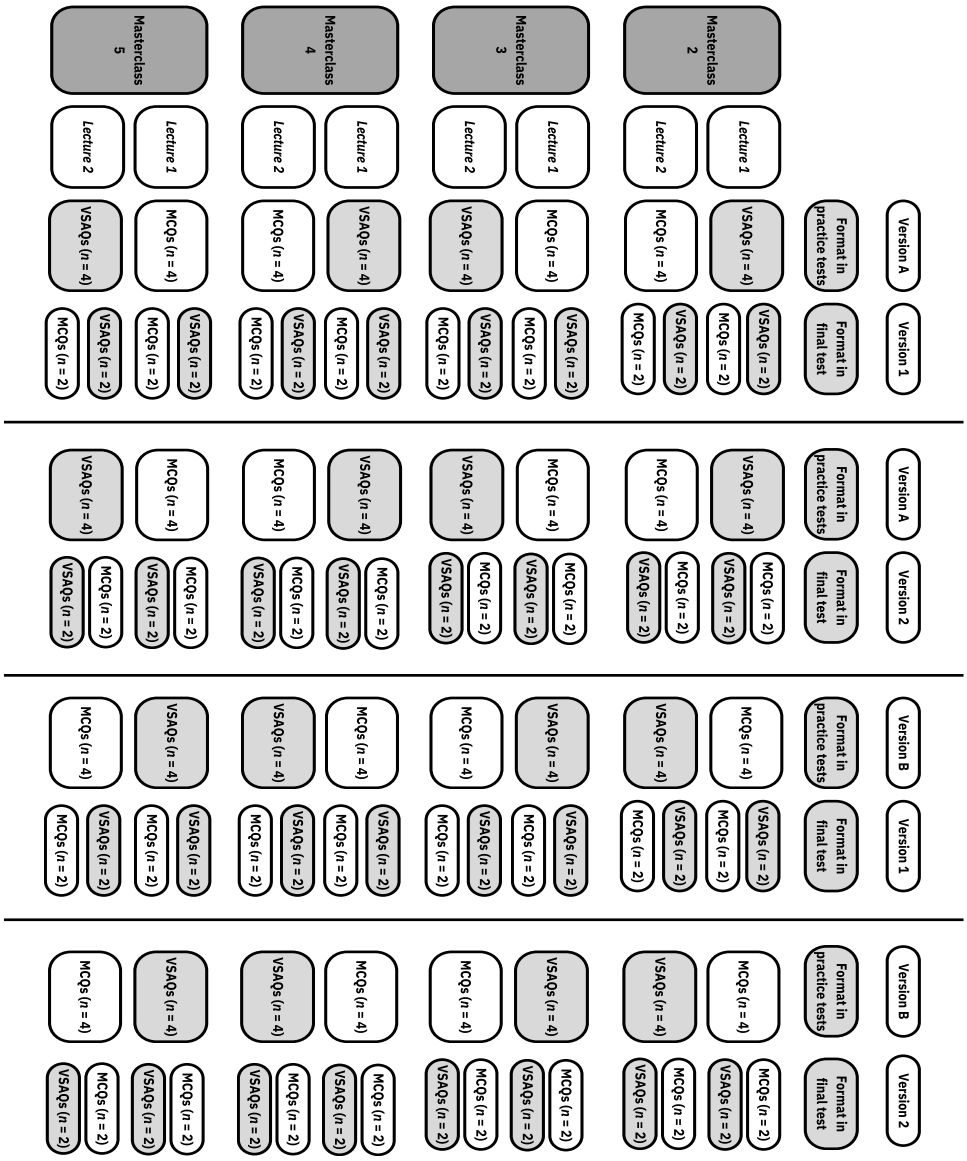
4. What type of diet is associated with an increased risk of mental health issues?
   A. Fermented food
   B. Highly processed food
   C. Herbs and spices
   D. Fat fish

Very short answer questions

You have to answer these questions with a very short answer (one to four words)

1. In about 40% of the complaints, no clear physical cause is found in the general practice. In what percentage of cases do such complaints become chronic?
2. Pain is an important protective mechanism of the brain. What is it trying to protect you from?
3. What is the first intervention for patients suspected of having SOLK/ALK?
4. There is a theory that suggests that the brain actively makes estimates about upcoming sensory input instead of passively registering it. This could play a role in the development of chronic pain complaints. What is this theory called?

Appendix 3 – Supplemental Figure 1



Supplemental Figure 1. Detailed study design with versions A and B (practice tests) and versions 1 and 2 (final test)

Appendix 4 – Questionnaire

Before the final test:

1. Have you prepared for the practice test? (yes or no). If yes, how? *<open answer>*
2. Have you prepared for the final test? (yes or no). If yes, how? *<open answer>*
3. Do you feel that you have better remembered the content of the lectures by taking the practice tests?
   - o Yes
   - o No
4. What percentage of the practice test about lecture 1 [title]; lecture 2 [title]; lecture 3 [title]; lecture 4 [title]; lecture 5 [title]; lecture 6 [title]; lecture 7 [title]; lecture 8 [title] do you think you have remembered?

After the final test:

5. How much of the material from this test do you think you will still know in 6 months? (*answer in percentage*)
6. Can we approach you to retake the same final test in 6 months?
   - o Yes
   - o No

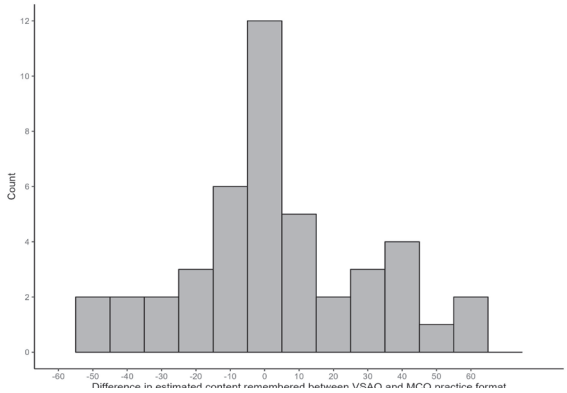
Appendix 5 – Supplemental Table 1

Supplemental Table 1. Mean percentages correctly answered questions on the final test for the different question formats on the practice test and final test.

Question format practice test	Question format final test	
	Multiple-choice question	Very short answer question
Multiple-choice question	64.5% (3.1)	49.7% (2.7)
Very short answer question	57.5% (3.1)	44.8% (3.3)

Standard deviations in parentheses.

Appendix 6 – Supplemental Figure 2



Supplemental Figure 2. The distribution of the difference in estimated content remembered practiced with VSAQs and MCQs. The percentages of lectures practiced with MCQs are subtracted from the percentages of lectures practiced with VSAQs.

Part I: Very short answer question

Chapter 5

Bridging assessment and clinical practice: The added value of very short answer questions in medical education

Elise V. van Wijk
Mario de Jonge
Floris M. van Blankenstein
Roemer J. Janse
Alexandra M.J. Langers

A version of this manuscript has been accepted in The Clinical Teacher (2025)

Introduction

Assessment plays a critical role in the career of medical doctors, influencing not only the preparedness of students for clinical practice but also shaping the ongoing development of clinical reasoning skills in future physicians. Not all healthcare professionals who are involved in assessing medical students or residents have received formal training in educational methodologies. For decades, medical education has relied predominantly on multiple-choice questions (MCQs) for assessment. While MCQs are praised for their reliability and ease of automated grading, they have limitations. Specifically, they often fail to adequately assess higher-order cognitive skills, such as clinical reasoning and problem-solving, and can promote superficial learning approaches that do not translate well into real-world patient care [1-3]. The concept of '*assessment drives learning*' highlights how the format and design of assessments influence the way students engage with learning material. Given the limitations of MCQs, there is a growing need for alternative question formats that better promote deep learning and better prepare students for the clinical challenges they will face [4]. Very short answer questions (VSAQs)—open-ended questions that require a concise response of 1–4 words—have been proposed as promising alternative. VSAQs may better promote deep learning and simulate real-world clinical reasoning, as students must actively retrieve and articulate answers without the cues present in MCQs. This article explores the potential of VSAQs to improve medical education and how they may ultimately contribute to enhanced patient care.

The purpose and impact of assessment in medical education

Assessment serves various purposes. The most familiar is summative assessment, measuring whether students have acquired sufficient knowledge to meet certain academic or professional standards. This *'assessment of learning*' results in grades or credits and is used to certify competence at various stages of medical training. However, there is a growing emphasis on the formative role, in which assessments guide students' learning and help them improve through feedback—known as 'assessment for learning' [5]. Practice tests, used as formative assessments, enable retrieval practice, where students actively recall information from memory rather than passively reviewing material. For example, students might self-assess by quizzing themselves on the steps for managing a patient with sepsis, rather than rereading lecture notes. Retrieval practice, also known as the testing effect, strengthens memory and promote long-term retention, which is crucial for applying core clinical knowledge in clinical practice [6]. Understanding the different functions of assessment is relevant not only for those involved in teaching and assessing medical students and residents, but also for all clinicians, as it underscores the importance of assessment as a tool for lifelong learning.

High-quality assessment requires carefully constructed questions that consistently measure the intended knowledge (reliability) and align with learning objectives (validity). This principle of *'constructive alignment'* [7] ensures assessments to support and enhance learning. However, the effectiveness of assessment also depends on the design of the question formats used, which can significantly impact both their validity and the level of understanding they aim to measure.

Limitations of the traditional assessment question format

Traditionally, written assessments in medical education mainly consist of MCQs. These questions offer high reliability, and ease of marking for large student cohorts [8]. However, their reliance on recognition-

based learning strategies enables students to pass exams through surface-level preparation, often by recognizing correct answers rather than deeply understanding the material [1, 3]. Additionally, cues in the question or answer options can enable students to deduce correct answers without fully grasping the subject matter. This reliance on recognition or cues diminishes the discriminative power of MCQs to distinguish between students who truly understand the material and those who do not [9, 10].

Exploring new question formats: the Very Short Answer Question

With the increased adoption of digital testing, other question formats that were once impractical due to time and resource constraints, such as the labor-intensive process of manual marking, have become more accessible. One promising alternative is the Very Short Answer Question (VSAQ) (*Figure 1*). Advances in digital grading have made the use of VSAQs more feasible, even for large-scale assessments. VSAQs are particularly well-suited to medical education because they combine the depth of open-ended questions with marking efficiency approaching that of MCQs.

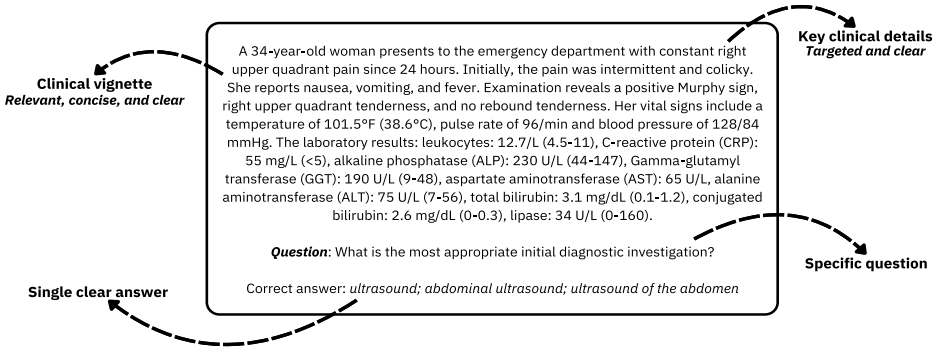


Figure 1. Example of a well-crafted very short answer question

VSAQs offer several advantages. Firstly, by requiring students to generate an answer rather than select one, they promote deeper cognitive processing [1, 11]. Moreover, VSAQs have higher discriminative power than MCQs, as they eliminate cues and minimize guessing. This enables more accurate differentiation between students who truly understand the material and those who do not [9, 10, 12]. VSAQs are also shown to be more reliable than MCQs, which might be explained by reducing the impact of partial knowledge, where students might guess correctly without a complete understanding [9, 10]. Additionally, VSAQ answers provide valuable insights into student misconceptions, cognitive processes, and errors. Unlike MCQs, which may mask gaps in understanding due to guessing or recognition of correct answers, VSAQs allow teachers to observe how students approach a clinical case and construct their answers. This richer understanding offers opportunities for targeted feedback to students and helps teachers identify areas where clinical teaching can be improved [1, 13, 14].

Another important advantage of VSAQs is their alignment with real-world clinical practice. In the clinical settings, doctors are seldomly presented with multiple-choice options to guide their decisions.

Instead, they must recall information from memory and apply it to new complex situations without external prompts [2]. VSAQs mimic this authentic cognitive process more closely than MCQs, providing a more relevant and effective assessment method for clinicians. Additionally, some topics cannot be effectively assessed using MCQs due to the difficulty in designing questions with plausible distractors. As a result, the questions risk focusing on testing less relevant, detailed factual knowledge rather than assessing core knowledge [13]. VSAQs help overcome this limitation, allowing for the assessment of essential knowledge without the need for other (incorrect) answer options. Furthermore, for teachers, VSAQs can simplify the question writing process. Creating plausible but incorrect distractors for MCQs is often challenging and time-consuming, whereas VSAQS eliminate this requirement, allowing for more efficient question development.

Despite their advantages, VSAQs also come with certain challenges. Writing effective VSAQs requires careful attention, as vague or ambiguous phrasing can create uncertainty for students, especially those accustomed to MCQs. Uncertainty often arises when students are unclear about the desired level of specificity in their responses. To address this, it is essential to provide clear, specific lead-ins for each question particularly when they are unsure about the desired level of specificity in their responses [9, 13]. For instance, the question *'What is the first-choice treatment?'* for a patient with hypertension is too broadly stated, leaving it unclear whether lifestyle measures or medication are being referred to. A clearer alternative would be '*What is the most appropriate pharmacological treatment?*' (also see *Figure 1* for an example). Furthermore, it is important to ensure that the knowledge being assessed is relevant and necessary to recall from memory — focusing on core clinical knowledge rather than peripheral details. Aligning the type of knowledge assessed with the most suitable question format is critical for creating meaningful and effective assessments.

Implementing VSAQs into the medical curriculum: strategies for success

Although VSAQs show promise, their adoption remains limited, particularly in medical education, where MCQs dominate. One concern is the time and effort required for grading. However, advances in digital testing systems have significantly reduced the burden of grading, enabling efficient grading even for large cohorts. Recent studies demonstrate that VSAQs can be graded with an average of two minutes per question for a cohort of approximately 350 medical students [9, 10]. Artificial intelligence advancements could further reduce grading times, potentially making VSAQs as efficient as MCQs [15]. Moreover, VSAQs eliminate the need to create plausible distractors—a particularly time-consuming and challenging aspect of MCQ development—giving teachers greater flexibility to focus on designing meaningful and targeted assessment questions.

Faculty resistance may also arise from a lack of familiarity with the VSAQ format. Workshop sessions can help overcome this barrier, equipping teachers with the skills needed to design effective VSAQs and integrate them into the curriculum. Similarly, students might initially perceive VSAQs as more challenging than MCQs. However, with appropriate preparation, practice tests, and guidance, they can recognize the long-term benefits of this question format. Simultaneously, these VSAQ practice tests can provide valuable feedback for identifying knowledge gaps and uncover misconceptions [13, 14]. Several tips have been published on how to implement VSAQs in the curriculum [13].

Conclusion

Assessment is a pivotal element of medical education, shaping both student learning and clinical preparedness. While MCQs have long dominated this space due to their reliability and practicality, they often fail to assess higher-order thinking and can encourage surface learning. VSAQs address these limitations by requiring active recall, encouraging active learning, motivate students to thoroughly understand the study material, and mirroring the clinical practice. VSAQs also offer teachers valuable insights into student misconceptions and cognitive processes, enabling targeted feedback and improved teaching strategies. Although concerns about grading time and unfamiliarity with the format exist, advances in digital testing and emerging technologies are making VSAQs increasingly practical for large-scale assessments. Additionally, VSAQs simplify question design by eliminating the need for plausible distractors, further enhancing their feasibility. For clinical doctors involved in teaching and assessment, VSAQs provide an effective way to foster meaningful learning and better prepare students for the clinical practice, ultimately enhancing both education and patient care.

References

1. Students' cognitive processes when answering two different formats of written question. Medical Teacher. 2021;43(11):1278-85.
2. 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.
3. 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.
4. Newble DI, Jaeger K. The effect of assessments and examinations on the learning of medical students. Medical Education. 1983;17(3):165-71.
5. Schuwirth LWT, Van der Vleuten CPM. Programmatic assessment: From assessment of learning to assessment for learning. Medical Teacher. 2011;33(6):478-85.
6. 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.
7. Biggs J. Enhancing teaching through constructive alignment. Higher Education. 1996;32(3):347-64.
8. Al-Rukban MO. Guidelines for the construction of multiple choice questions tests. Journal of Family & Community Medicine. 2006;13(3):125-33.
9. van Wijk EV, Janse RJ, Ruijter BN, Rohling JHT, van der Kraan J, Crobach S, et al. Use of very short answer questions compared to multiple choice questions in undergraduate medical students: An external validation study. PloS One. 2023;18(7):e0288558.
10. Sam AH, Field SM, Collares CF, van der Vleuten CPM, Wass VJ, Melville C, et al. Very-short-answer questions: reliability, discrimination and acceptability. Medical Education. 2018;52(4):447-55.
11. 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
12. Schuwirth LWT, Vleuten CPM, Donkers HHLM. A closer look at cueing effects in multiple-choice questions. Medical Education. 1996;30(1):44-9.
13. 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.
14. Putt O, Westacott R, Sam AH, Gurnell M, Brown CA. Using very short answer errors to guide teaching. The Clinical Teacher. 2022;19(2):100-5.
15. Grévisse C. LLM-based automatic short answer grading in undergraduate medical education. BMC Medical Education. 2024;24(1).

Part II: Computer adaptive progress testing

Chapter 6

Computer adaptive vs. non-adaptive medical progress testing: Feasibility, test performance, and student experiences

Elise V. van Wijk
Jeroen Donkers
Peter C.J. de Laat
Ariadne A. Meiboom
Bram Jacobs
Jan Hindrik Ravesloot
René A. Tio
Cees P.M. van der Vleuten
Alexandra M.J. Langers
André J.A. Bremers

Perspectives on Medical Education. 2024;13(1):406-216.
DOI: 10.5334/pme.1345

Abstract

Background: Computerized adaptive testing tailors test items to students' abilities by adapting difficulty level . This more efficient, and reliable assessment form may provide advantages over a conventional medical progress test (PT). Prior to our study, a direct comparison of students' performance on a computer adaptive progress test (CA-PT) and a conventional PT, which is crucial for nationwide implementation of the CA-PT, was missing. Therefore, we assessed the correlation between CA-PT and conventional PT test performance and explored the feasibility and student experiences of CA-PT in a large medical cohort.

Methods: In this cross-over study medical students (*n*=1432) of three Dutch medical schools participated in both a conventional PT and CA-PT. They were stratified to start with either a conventional PT or CA-PT to determine test performance. Student motivation, engagement and experiences were assessed by questionnaires in students from seven Dutch medical schools. Parallel-forms reliability was assessed using the Pearson correlation coefficient.

Results: A strong correlation was found (0.834) between conventional PT and CA-PT test performance. The CA-PT was administered without system performance issues and was completed in a median time of 83 minutes (67-102 minutes). Questionnaire response rate was 31.7% (526/1658). Despite a higher experienced difficulty, most students reported persistence, adequate task management and good focus during the CA-PT.

Conclusions: CA-PT provides a reliable estimation of students' ability level in less time than a conventional non-adaptive PT and is feasible in students throughout the entire medical curriculum. Despite the strong correlation between PT scores, students found the CA-PT more challenging.

Introduction

In the mid-1970s, Maastricht medical school introduced the progress test (PT) to align the assessment system with the rationale of the innovative instructional method of problem-based learning. This initiative aimed to mitigate the test-directed learning stimulated by end-of-unit assessments [1]. By introducing this comprehensive test, which aims to assess the end objectives of the medical curriculum, specific test preparation was discouraged. Its longitudinal design together with the feedback enhances the educational impact, by fostering long-term learning of functional knowledge [2-7]. To ensure a valid and reliable test content, the Dutch PT uses a blueprint containing a prescribed distribution of items across medical classifications and disciplines [8]. When the study was conducted the PT was implemented in several countries as a paper- or computer-based test, consisting primarily of multiple-choice questions (MCQs) [3]. Today, students of all Dutch medical schools participate in a national PT [9]. As in this fixed linear test format the knowledge level of individual students is not considered, the test contains items at a distance from the students' ability, which is likely to lower the test's reliability; an important criterium for good assessment [7, 10-12]. Furthermore, with the increasing number of participating medical schools during the past years, the simultaneous administration of progress tests to all students nationwide has become a logistical and costly challenge, limiting the feasibility of the test [7, 12].

Computerized adaptive testing (CAT) is a form of digital assessment that delivers a more tailored test to individual students by adapting the questions to the examinee's ability level using a pre-determined algorithm [13]. Usually, CAT adapts the difficulty level of the questions to the performance of the student during the exam. However, there are also other forms of computer adaptive testing available, each with their own assumptions, merits, and limitations. Some examples are multidimensional CAT [14], content-based CAT [15], testlet-based CAT [16] , and tree-based CAT [17]. Our focus is on a CAT that is based on plain Item Response Theory (IRT), which is a cornerstone of modern test theory. Unlike the Classical Test Theory (CTT), which is the underlying theory of the conventional Dutch PT's fixed linear test format, it does not assume that each question (or item) is equally difficult. Instead, it uses mathematical models to estimate the underlying ability level ('theta') of the test-taker based on their responses to different items. Each item is characterized by parameters that reflect its difficulty and discrimination, which allows for the creation of tests that are tailored to the test-takers ability level [18]. CTT on the other hand assumes that the observed total test score equals the actual ability level of the test-taker ('true test score') with an identical measurement error for all scores. These assumptions can lead to less precise estimates of a test-taker's ability [19]. As such, the CAT provides a more efficient test by reducing test length on average by 50% while preserving or even improving the reliability of the test [10, 11, 20, 21]. Moreover, the Online Adaptive International Progress Test (OAIPT) project showed that the adaptive test was well-accepted by students and might improve motivation and engagement, which was also demonstrated earlier in elementary and high school students [10, 22, 23]. Effective development and feasibility of implementing a computer adaptive PT (CA-PT) in medical education across several European countries has been demonstrated before [11]. Simultaneous test administration, to prevent fraud by sharing exam information, is no longer required with the use of an online tailored test, reducing logistical issues, and improving feasibility.

Considering the benefits of CAT, it has been considered as a promising alternative for the CTT-based fixed linear test format of the conventional Dutch PT. While several studies have demonstrated strong correlations between fixed-length short forms and CAT in patient-outcome measurements [24-26], there is a lack of research comparing test performance on a linear-fixed PT with a CA-PT; a comprehensive, longitudinal test that adapts to the ability level of the student, administered to students at various curricular ages. A direct comparison between a CA-PT and conventional PT, in the same cohort of students and in an authentic setting, has yet to be conducted. This comparison is a necessary step towards the ambitious goal of implementing the CA-PT at a national level across all medical schools. Therefore, we aimed to 1) evaluate the correlation between test performance on a CA-PT and a conventional PT, and 2) assess the feasibility and student experiences of a CA-PT in a large cohort of Dutch medical students who were offered both a conventional PT and CA-PT.

Methods

Setting

The Dutch interuniversity medical PT is a longitudinal comprehensive test that covers the whole medical curriculum. In the Netherlands, the medical curriculum consists of a preclinical Bachelor and clinical Master phase, both with an average duration of three years each. The preclinical phase is made up of a variety of theoretical courses. Each of these courses is assessed by a summative assessment to evaluate a student's knowledge. The clinical phase is primarily composed of clinical rotations, which are separately or collectively evaluated by a summative pass/fail decision based on feedback from supervisors. The learning outcomes of the medical curriculum are described in a Framework for Undergraduate Medical Students, and are identical for all medical schools [27]. At the time of the study, seven of the eight Dutch medical schools participated in the PT. Throughout the six-year medical program, the PT is administered four times each academic year (September, December, February, and May), resulting in a total of 24 test moments for an individual student. The longitudinal design provides insights into a student's functional knowledge development over time in relation to peer medical students across the Netherlands. The conventional non-adaptive PT consists of 200 MCQs and is identical for all participating students. The questions are selected from an item bank based on a blueprint with a predetermined distribution covering all relevant medical disciplines and categories *(Appendix 1 – Supplemental Table 1*). The MCQs include a '*I don't know*' option symbolized by a question mark. Selection of this option results in a neutral score of zero points. An incorrect answer, on the other hand, incurs a penalty that results in a negative score. This so called formula scoring method encourages students to recognize their knowledge gaps and discourages random guessing [28]. The severity of the penalty of an incorrect answer is determined by the number of answer options. For instance, an incorrect answer in a MCQ with three options leads to a deduction of 1/3 points. This ensures that the penalty is proportional to the probability of guessing the correct answer. The final score is computed as the sum of the scores per MCQ and is expressed as a percentage of the maximum attainable score, and is translated into '*Good*', '*Pass*', or '*Fail*', based on the mean and standard deviation of the complete student cohort in the same test moment as a relative standard. Progress in academic years goes along with increased passing scores of the PT. At the end of each academic year, the results of the four formative progress tests are combined into a summative decision (fail, pass, or good) [9].

Development of the question bank

At the time of the study, the CA-PT item bank consisted of 3400 calibrated questions. These questions originate from 30 previous linear progress tests, spanning a period of 7.5 year. All questions were reviewed according to a rigorous peer-review process to determine if they were still correct and up-to-date before adding them to the item bank. Using the answer data from these historical 30 tests, we calibrated these questions following a Rasch model, a widely used IRT approach, to obtain their difficulties [29]. Question pairs assessing the same topic in a textual similar way, and conflicting questions were classified as enemy items, meaning that the system prevents usage of these questions in a single test. Before the questions had been used in a PT, they received a label for '*Category*' and '*Discipline*', which places them in individual cells of the blueprint (*Supplemental Table 1)*.

Question selection in the CA-PT

The CA-PT consists of 135 MCQs without a question mark option; 120 calibrated questions, and 15 non-adaptive pretest questions. Every student receives questions according to the PT blueprint (*Appendix 1 – Supplemental Table 1*).The decision to use a fixed number of 120 questions was driven by our objective to reduce the overall length of the PT while still sufficiently covering the blueprint. We use a fixed-length CAT to provide a similar test experience for all students. The pretest questions are seed items (newly written or revised questions), randomly distributed throughout the CA-PT, are included for calibration, and do not contribute to the test result. After calibration, these new questions are added to the item bank for subsequent use. Prior to the adaptive phase of the CA-PT (i.e., 114 questions), six non-adaptive calibrated starter questions are administered to make a first estimation of the student's ability level. The average difficulty level of these six questions together is zero, and the questions count for the test result. Due to the adaptive nature of the CA-PT, navigation is only unidirectional, whereas in the conventional PT students had the possibility to review previously answered questions during the test and change their answer if desired. The score of the CA-PT is the estimated ability level based on the answers on the 120 calibrated questions selected by the algorithm combined with the item difficulty of the questions [30].

Study design and data collection

In this cross-over study students participated in both a conventional PT and CA-PT in May 2022, which was the last PT of the academic year 2021-2022. The conventional PT was mandatory for all students, and participation in the CA-PT was voluntary. To encourage students to perform at their best in both tests, the highest outcome was taken into account for their study progress. Students were stratified to start with either a conventional PT (PT*first*) or a CA-PT (CA-PT*first*) based on a fixed availability of the timeslots for each test moment. The conventional PT was administered as a paper-based test and the CA-PT as a digital test in TestVision®. Both PTs were administered in an exam hall with supervision. The conventional PT was administered to all students on the same day, during the same time slot. The allotted time to complete the PT was 240 minutes for the conventional PT, and 180 minutes for the CA-PT. The time interval between the conventional PT and the CA-PT for an individual student was seven days or less. The test results were communicated to students by email after two weeks for the conventional test, and after five weeks for the CA-PT.

On completion of the CA-PT, digital questionnaires were administered to gain insights into the student experiences (*Appendix 2 – Questionnaire*). All students had previous experience with the paper-based

PT. At the time of administration of the questionnaire, students were unaware of their test results. Items 1-11 of the questionnaire were derived from the Short Motivation and Engagement Scale (six items on positive, and five items on negative test-relevant motivation, and engagement), adapted to our context and translated to Dutch [31]. Items 12-15 assessed the subjective experience of the CA-PT in comparison to the conventional PT, and were based on the questionnaire used in the study by Martin & Lazendic [22]. Five out of the seven items were found relevant to include in our questionnaire.

Participants

Students from all participating medical schools were offered the opportunity to participate in a CA-PT of May 2022. In three of the participating medical schools (MS1, MS2, and MS3) the CA-PT could be offered to all students under full study conditions. Due to logistic issues and/or a lack of approval by the local board of examiners, students from the other four medical schools were not able to participate in the study, although some students had the opportunity to try-out the CA-PT without the result being taken into account. Students who participated in both a CA-PT and conventional PT in MS1, MS2, and MS3 were included for analyses regarding test performance. Regarding feasibility of CA-PT administration, and student experiences, we analyzed the data of all participants of the seven medical schools. The PT in May 2022 (the fourth PT of the academic year) entailed test moments 4 (year 1), 8 (year 2), and 12 (year 3) for the bachelor students and 13 to 24 for the master students, as master students enter the master phase at different timepoints throughout the year. For master students in Erasmus MC, this was only test moment 13 to 16, as the PT was introduced there in September 2021 for the master.

Information materials about the CA-PT were developed on a national level, and used by all medical schools. There were short animations about the CA-PT (see for example *https://www.youtube.com/watch?v=xjwHLhXhIho*), written information and frequently asked questions on the Dutch PT-website [32]. A national webinar for students was organized and recorded for later use. Furthermore, individual medical schools communicated identical information with their students via their local communication systems and/or organized (web)lectures.

Data analysis

To assess possible differences in PT scores between PT*first*, and CA-PT*first* of the three participating medical schools we used z-scores, and an unpaired t-test. We also compared the z-scores of students who participated in our study with the z-scores of students who only participated in the conventional PT. The z-scores were calculated for the conventional PT, and CA-PT relative to all students in the same test moment group, providing a level of each student relative to their peers. Effect sizes were determined by the Cohen's *d* coefficient. The Pearson correlation coefficient was utilized to evaluate the correlation between the total score on the conventional PT, and the theta (ability level) [33] on the CA-PT across both tests. The total score of the conventional PT was selected for this analysis, as this includes the question-mark option in the score. This question-mark option, and thereby the decision to answer a question or not, is an essential part of the conventional PT. Consequently, this approach provided the most reliable and authentic method for comparing the different PT formats. Characteristics of responders to the questionnaires are presented as mean (standard deviation), or median (interquartile range) depending on their distribution. Categorical variables are presented as number (proportion). All statistical analyses were performed using R version 4.1.0 (R Foundation for Statistical Computing, Vienna, Austria).

Ethical approval

The approval to conduct this study was granted by the Ethical Review Board of the Netherlands Association for Medical Education (NVMO): NERB/2023.4.6. Participation in the CA-PT was voluntary, and all students received verbal and written information prior to the study. Upon initiation of the CA-PT, students provided informed consent.

Results

In total 1432 students (647 bachelor, 785 master) from MS1, MS2, and MS3 were included in our analysis regarding student performance. In the other medical schools, a total of 226 students took part in the CA-PT, but their test results were not taken into account in the performance calculations as the study conditions were not met. Of the 1658 participating students in all medical schools, 526 students (response rate 31.7%) completed the questionnaire (*Figure 1*).

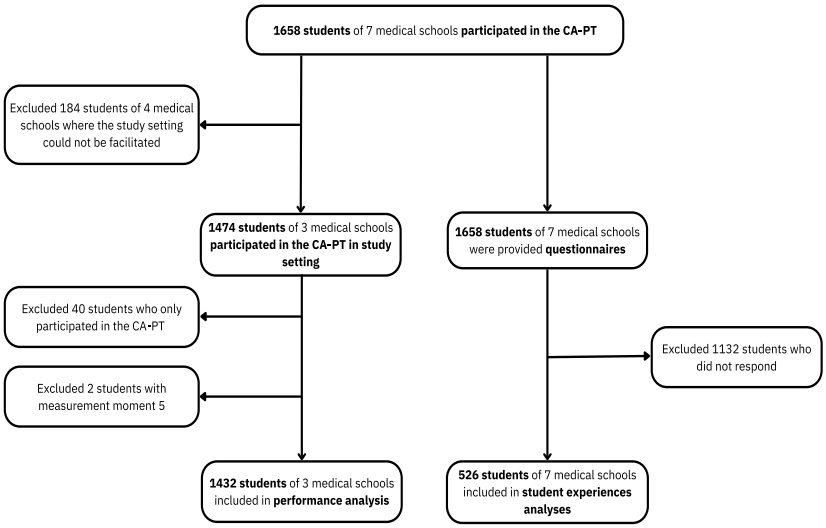


Figure 1. Flowchart of participants in questionnaire and test performance analyses

Test performance

Students in the PT*first*-group (n=797; mean (M)=0.406, SD=1.06) performed slightly better on the conventional PT compared to students in the CA-PT*first*-group (n=635; M=0.24, SD=1.03; t(1373) = 3.08; p=0.002; Cohen's d= 0.16). No difference was found in performance on the CA-PT between both groups (t(1345)=-1.0324, p=0.302). Within the three participating medical schools there was a small but significant difference between the conventional PT scores of students who participated in both a conventional PT and CA-PT, and students who participated only in a conventional PT in MS1 (M=0.38, 0.19, SD=1.09, 0.98; t(1444)=3.49, p<0.001; Cohen's d=0.18), and MS2 (M=0.28, 0.16; SD=0.99, 0.93; t(738)=2.24; p=0.025; Cohen's d=0.13), but not in MS3 (t(551)=0.59, p=0.551). The parallel-forms reliability, i.e. the correlation between the total score of the conventional PT, and the theta of the CA-PT was 0.834.

After adjustment for the differences in PT score between PT*first* and CA-PT*first* the correlation becomes slightly less: 0.832. The correlation was moderate within each year group: 0.506 (Y1; *n*=253), 0.675 (Y2; *n*=211), 0.754 (Y3; *n*=183), 0.733 (Y4; *n*=414), 0.708 (Y5; *n*=164), and 0.673 (Y6; *n*=207) (*Figure 2*).

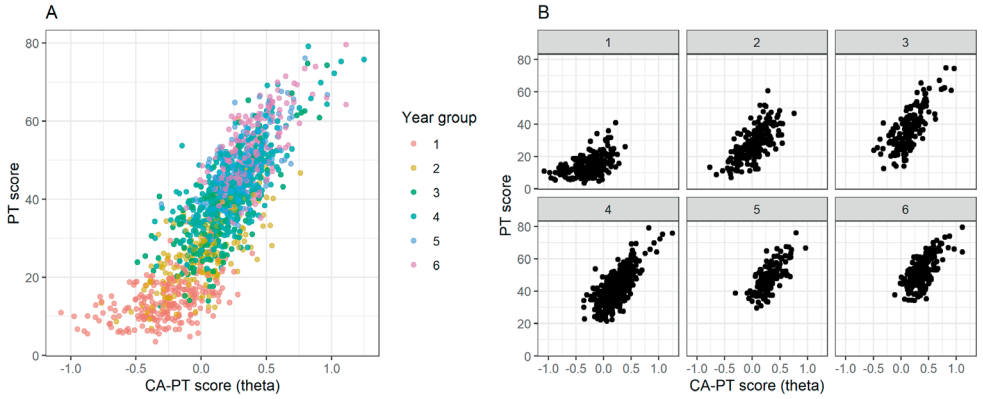


Figure 2. The relationship between the z-scores on the conventional PT (y-axis) and the theta on the CA-PT (x-axis) for ***(A)*** year 1 to 6 and ***(B)*** each year separately.

Feasibility of the CA-PT

Ninety percent of students finished the CA-PT within two hours (median: 83 minutes; IQR: 67-102 minutes). There were no performance issues with the digital assessment system. The algorithm was able to select questions of an appropriate level, defined as a difference of less than 0.1 between the estimated ability of the student, and pre-calibrated level of difficulty of the question, in more than 99% of the questions.

Student motivation, engagement and experiences

Of the 526 responders to the questionnaire, 451 students were from MS1, 2, and 3. The responders had a mean age of 22.8 (3.0) years, and 74.1% were female. The median test moment was 15 (IQR: 8-19). Eighty-four percent of the students agreed that they persisted even when the CA-PT was challenging or difficult. Most students did not want to receive a bad grade for this exam (77%), made good use of their time during the CA-PT (71%), and were focused on understanding the questions (71%). The majority of students were not anxious (63%) or felt like giving up during the CA-PT (61%). Almost 80% of the students experienced the CA-PT as more difficult compared to the conventional PT. A total of 76% of students did not think they performed better on the CA-PT, and 24% of students thought that the CA-PT was better adjusted to their level (Figure 3). For approximately 90% of the students, the provided information on CAT was clear and they knew what to expect from the CA-PT. In response to the open question regarding their experience with the CA-PT (*n*=422) the majority of comments were about: 1) missing the option to go back to the previous question (*n*=112), 2) missing the question mark option (*n*=87) and 3) it being more difficult to predict their performance level, leading to higher levels of insecurity and nervosity, and/or decreased motivation (*n*=87).

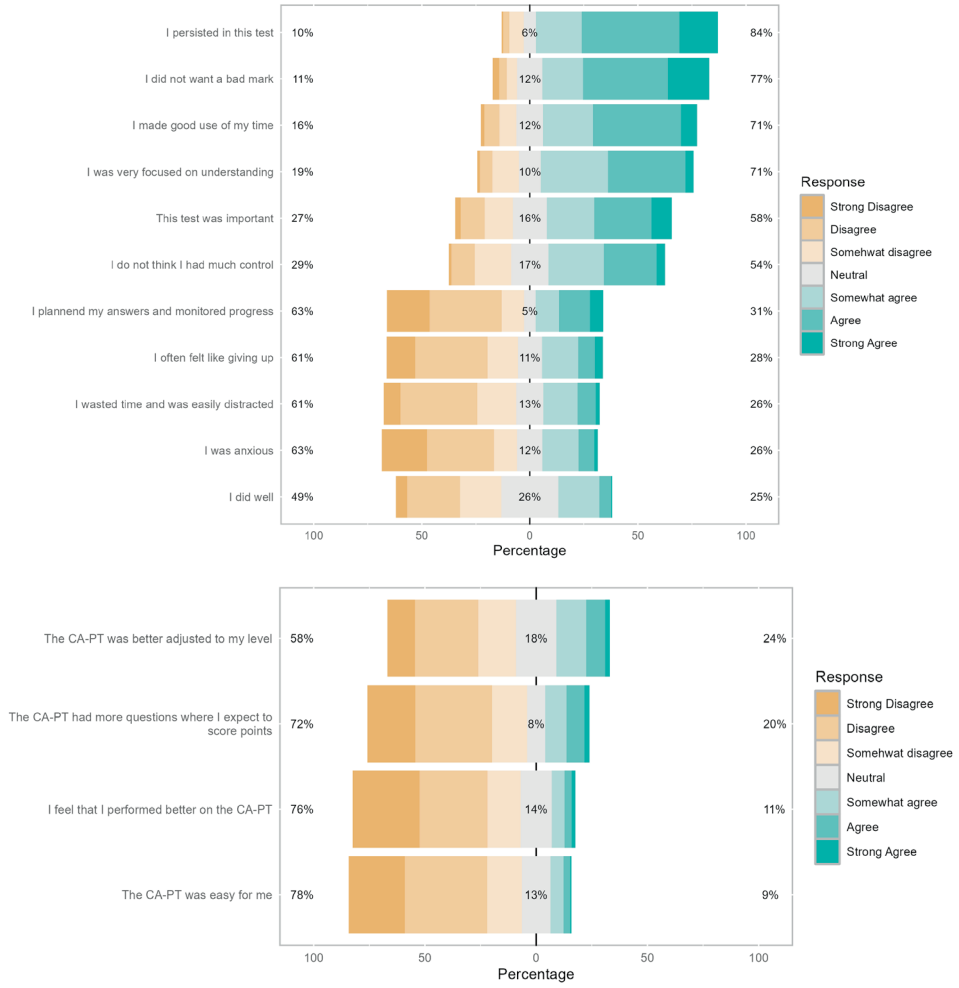


Figure 3. Distribution of answers to the questionnaire items ***(A)*** 1 to 11 on motivation and engagement in the computer adaptive progress test; ***(B)*** 12 to 15 comparing the conventional progress test with the computer adaptive progress test. I persisted in this test = "*I persisted in this test even when it was challenging or difficult*"

Discussion

To our knowledge, this study is the first to investigate test performance on a CA-PT compared to a conventional PT in an authentic setting with a large cohort of medical students at different study stages. A cross-over design was used in which all participating students were offered both a conventional PT and a CA-PT at a single timepoint and in an authentic examination setting. We found a strong correlation between test performance on the conventional PT and CA-PT. The CA-PT was administered without system performance issues, most students finished the CA-PT within two hours and students were motivated to perform well, despite the feeling that the CA-PT was more difficult.

The overall strong correlation between test performance on the conventional PT and CA-PT demonstrates that the CA-PT is able to reliably determine a students' aptitude after a significantly shorter test. However, if we look at the correlation within the different year groups, the correlation was weaker in first-year students (r = 0.506) . This may be explained by the fact that they had to answer all questions in the CA-PT and could not decide to use the question mark option in case they did not know the answer. More frequent use of the question mark in first year students lowers the total amount of answered questions and thereby the reliability of the conventional PT. In contrast, the larger amount of answered questions in the CA-PT ensures a more accurate, reliable score calculation, with a possibly larger variance in scores, which might explain the weaker correlation with the conventional PT in these students. Test reliability of the CA-PT is shown to be high for students across the full spectrum of ability, and thereby improves test reliability and quality especially for students in the first years of their study [11].

Overall, students were motivated, and engaged to perform well in the CA-PT. Although the students perceived the CA-PT as more difficult compared to the conventional PT, this was not reflected by poorer test performance. With respect to their attitude towards the CA-PT, our questionnaire data suggest that most students were persistent, had a mastery orientation, and adequate task management in the CA-PT. Additionally, the majority of students did not experience negative test-related motivation, and engagement, such as anxiety, self-handicapping ("*During this test I wasted time and was easily distracted"*), and disengagement (*"I often felt like giving up in this test"*). Our findings align with improved motivation for learning, and engagement with the test in the OAIPT project [23], and in elementary and secondary school students [22]. In contrast to this study [22], we did not find the specific factors self-efficacy and anxiety to be increased, although the open question reveals higher levels of insecurity and nervosity regarding performance level than answers to the closed questions suggest. Lower self-efficacy and increased insecurity may both be related to the degree of perceived control and the feeling that the items are well-matched to their performance level, as these factors are suggested to promote self-efficacy and diminish anxiety in CAT [34-36]. Nevertheless, these negative feelings were not accompanied by reduced motivation and engagement, which might be related to the fact that students felt challenged, well informed, knew what to expect, and were provided two opportunities to perform on the PT [22, 37].

Strengths and limitations

This multi-center study is the first to assess both test performance and test experience of a CA-PT in an authentic setting with medical students at different stages throughout the entire medical curriculum. The cross-over design and the short interval between the tests enabled us to compare performance within students at a given point in time, while the possible benefits for the students (best outcome counts for study credits) stimulated optimal test effort in both tests. However, the difference in delivery between the test formats, paper-based versus digital, might have influenced student performance depending on their preferences, though our experiences during the COVID-19 pandemic suggested that the effect on performance using different delivery formats is minimal (unpublished data). The difference in feeling of success, or certainty about their performance between the test formats might have had a psychological impact that differs between students. In the conventional PT, students usually experience a sense of how well they performed, derived from the proportion of items that they answered with certainty. In the CA-PT this sense is absent, as the number of wrong answers is approximately 50% for each individual. Our study sample was representative for all students participating in the conventional PT within the three medical

schools where the study setting was facilitated, despite a slight overrepresentation of better performing students in two of the schools. Although the PT*first* and CA-PT*first* group were comparable in their performance on the CA-PT, students in the PT*first*-group performed slightly better on the conventional PT. Because students experienced decreased accuracy in estimating their performance on the CA-PT, or because they could review the questions of the conventional PT with the answer key directly afterwards, students in the PT*first*-group might have experienced less pressure to perform at their best in the CA-PT. Regardless, the effect of this group difference on the correlation was negligible (0.834 to 0.832). The study setting could only be facilitated in three of the seven medical centers. Still, the number of participants was large enough to leave our analysis of test performance uncompromised. Finally, two-thirds of the students who participated in the CA-PT did not return the questionnaire, which might have caused a bias regarding students' opinion.

Implications and future research

Taken together, our results support a broader application of the CA-PT in medical progress testing. As motivation, engagement and subjective test experience may affect students' willingness to put effort in the test, and thereby influence their performance, it is relevant to shed light on these aspects [22, 38]. Our finding that most students experienced the CA-PT as more difficult, and felt insecure about their performance, is important to take into consideration when preparing students for this new test format. Also, the responses to the open questions indicate that students find it difficult to switch to a new testing format, emphasizing the need for clear information, and practice opportunities. An interesting direction of future research could be the exploration of test performance, and student experiences over a longer period, as students continue getting accustomed to this testing format.

Conclusion

In conclusion, this study shows that a CA-PT provides a reliable estimation of the students' aptitude with a reduced test length in medical students. Students were motivated and engaged to perform well on the CA-PT, despite experiencing it as a more difficult test. Therefore, the implementation of a CA-PT in a wider context seems justified.

References

1. Vleuten CPMVD, Verwijnen GM, Wijnen WHFW. Fifteen years of experience with progress testing in a problem-based learning curriculum. Medical Teacher. 1996;18(2):103-9.
2. 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.
3. Dion V, St-Onge C, Bartman I, Touchie C, Pugh D. Written-Based Progress Testing: A Scoping Review. Academic Medicine. 2022;97(5):747.
4. 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.
5. 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.
6. Pugh D, Desjardins I, Eva K. How do formative objective structured clinical examinations drive learning? Analysis of residents' perceptions. Medical Teacher. 2018;40(1):45-52.
7. 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).
8. 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.
9. Dutch experience. Perspectives on Medical Education. 2016;5(1):51-5.
10. Collares CF, Cecilio-Fernandes D. When I say ... computerised adaptive testing. Medical Education. 2019;53(2):115-6.
11. 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.
12. Norcini J, Anderson B, Bollela V, Burch V, Joao Costa M, Duvivier R, et al. Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. Medical Teacher. 2011;33(3).
13. Chang H-H. Psychometrics behind Computerized Adaptive Testing. Psychometrika. 2015;80(1):1-20.
14. Wang C, Weiss DJ, Su S, Suen KY, Basford J, Cheville A. Multidimensional Computerized Adaptive Testing: A Potential Path Toward the Efficient and Precise Assessment of Applied Cognition, Daily Activity, and Mobility for Hospitalized Patients. Archives of physical medicine and rehabilitation. 2022;103(5).
15. Burr SA, Gale T, Kisielewska J, Millin P, Pêgo JM, Pinter G, et al. A narrative review of adaptive testing and its application to medical education. MedEdPublish. 2023;13.
16. Frey A, Seitz N-N, Brandt S. Frontiers | Testlet-Based Multidimensional Adaptive Testing. Frontiers in Psychology. 2016;7.
17. Delgado-Gómez D, Laria C. J, Ruiz-Hernández D. Computerized adaptive test and decision trees: A unifying approach. Expert Systems with Applications. 2019;117.
18. Downing SM. Item response theory: applications of modern test theory in medical education. Medical Education. 2003;37(8):739-45.
19. Traub RE. Classical Test Theory in Historical Perspective. Educational Measurement: Issues and Practice. 1997;16(4):8-14.
20. Tian J-q, Miao D-m, Zhu X, Gong J-j. An Introduction to the Computerized Adaptive Testing. 200.
21. 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.
22. Martin AJ, Lazendic G. Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. Journal of Educational Psychology. 2018;110(1):27-45.
23. 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.
24. Amtmann D, Bamer AM, Kim J, Bocell F, Chung H, Park R, et al. A comparison of computerized adaptive testing and fixed-length short forms for the Prosthetic Limb Users Survey of Mobility (PLUS-MTM). Prosthetics and Orthotics International. 2018;42(5):476.
25. Gibbons RD, Weiss DJ, Kupfer DJ, Frank E, Fagiolini A, Grochocinski VJ, et al. Using Computerized Adaptive Testing to Reduce the Burden of Mental Health Assessment. Psychiatric services. 2008;59(4):361-8.
26. Choi SW, Reise SP, Pilkonis PA, Hays RD, Cella D. Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation. 2010;19(1):125-36.
27. Framework for Undergraduate Medical Education 2021 [updated 2021-08-20; cited July 2023]. Available from: https://www.nfu.nl/en/themes/professional-future/medicine-programmes/framework-undergraduate-medical-education.

28. Lord FM. Formula scoring and number-right scoring. Journal of Educational Measurement. 1975;12(1):7-11.
29. Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests: MESA Press, 5835 S; 1993.
30. Warm TA. Weighted likelihood estimation of ability in item response theory. Psychometrika. 1989;54(3):427-50.
31. Martin AJ. Motivation and Engagement Across the Academic Life Span: A Developmental Construct Validity Study of Elementary School, High School, and University/College Students. Educational and Psychological Measurement. 2009;69(5):794-824.
32. iVTG website. Available from: https://ivtg.nl/nl/.
33. Hambleton RK,. Fundamentals of Item Response Theory. SAGE. 1991.
34. Parshall CG, Spray JA, Kalohn JC, Davey T. Practical Considerations in Computer-Based Testing. Springer; 2002.
35. Pitkin AK, Vispoel WP. Differences Between Self-Adapted and Computerized Adaptive Tests: A Meta-Analysis. Journal of Educational Measurement. 2001;38(3):235-47.
36. Colwell NM. Test Anxiety, Computer -Adaptive Testing, and the Common Core. Journal of Education and Training Studies. 2013;1(2):50-60.
37. Ortner TM, Caspers J. Consequences of Test Anxiety on Adaptive Versus Fixed Item Testing. European Journal of Psychological Assessment. 2011;27(3):157-63.
38. Wise SL. Effort analysis: Individual score validation of achievement test data. Applied Measurement in Education. 2015;28:237-52.

Appendix

Appendix 1 – Supplemental Table 1

Supplemental Table 1A. Blueprint used in the conventional progress test.

Discipline	Number of questions
Anatomy	13
Biochemistry/Genetics/Histology/Molecular Cell Biology	18
Surgery	17
Dermatology/Ear, Nose, Throat/Ophthalmology	14
Epidemiology/Statistics	8
Pharmacology	9
Physiology	11
Geriatrics	8
Gynecology/Obstetrics	7
General Practice	20
Internal Medicine	26
Pediatrics	12
Metamedics	5
Neurology	7
Pathology/Immunology/Microbiology	10
Psychiatry/Psychology	12
Social Medicine	3
Total	200

Supplemental Table 1B. Blueprint used in the computer adaptive progress test.

Discipline	Number of questions
Anatomy	7;8
Biochemistry/Genetics/Histology/Molecular Cell Biology	10;11
Surgery	10;11
Dermatology/Ear, Nose, Throat/Ophthalmology	8;9
Epidemiology/Statistics	4;5
Pharmacology	5;6
Physiology	6-7
Geriatrics	4-5
Gynecology/Obstetrics	4-5
General Practice	12
Internal Medicine	15-16
Pediatrics	7-8
Metamedics	3
Neurology	4-5
Pathology/Immunology/Microbiology	6
Psychiatry/Psychology	7-8
Social Medicine	1-2
Total	120

Appendix 2 – Questionnaire

1. At which university do you study?
    - o Amsterdam – University of Amsterdam
    - o Amsterdam – Free University
    - o Leiden University Medical Center
    - o University Medical Center Groningen
    - o Maastricht University Medical Center
    - o Radboud University Medical Center Nijmegen
    - o Erasmus Medical Center Rotterdam
2. What is your student number?
3. I performed in the conventional progress test (PT) on the 25th of May 2022.
    - o Yes
    - o No

The following questions concern the computer adaptive PT. *The answers are given on a 7-point Likert Scale (1 = strongly disagree, 2 = disagree, 3 = somewhat disagree, 4 = either agree or disagree, 5 = somewhat agree, 6 = agree, 7 = strongly agree).*

Motivation and Engagement

1. I did well in this test.
2. In this test I was very focused on understanding the questions and tasks.
3. This test was important.
4. I persisted in this test even when it was challenging or difficult.
5. In this test, I planned my answers and monitored my progress.
6. In this test I made good use of my time.
7. I was anxious in this test.
8. In this test I did not want to get a bad mark.
9. I do not think I had much control over how well I did in this test.
10. During this test I wasted time and was easily distracted.
11. I often felt like giving up in this test.

In the following questions we ask you to compare the computer adaptive PT to the earlier conventional progress tests. *The answers are given on a 7-point Likert Scale (1 = strongly disagree, 2 = disagree, 3 = somewhat disagree, 4 = either agree or disagree, 5 = somewhat agree, 6 = agree, 7 = strongly agree).*

Subjective experience of the computer adaptive PT

12. In comparison to the conventional PT, the computer adaptive PT was easy for me.
13. In comparison to the conventional PT, the computer adaptive PT was better adjusted to my level.
14. In comparison to the conventional PT, there were more questions in the computer adaptive PT where I expect to score points.
15. In comparison to the conventional PT, I have the feeling I performed better on the computer adaptive PT.

Part II: Computer adaptive progress testing

Chapter 7

The effect of the question mark option in progress testing: A large-scale longitudinal study

Elise V. van Wijk
Jeroen Donkers
Peter C.J. de Laat
Ariadne A. Meiboom
Bram Jacobs
Jan Hindrik Ravesloot
René A. Tio
Frederike M.M. Oud
Jeroen P. Kooman
André J.A. Bremers
Alexandra M.J. Langers

Resubmitted after revisions to Perspectives on Medical Education (2025)

Abstract

Background: Formula scoring, widely used in medical progress tests (PT), includes a question mark option to discourage guessing, but this feature may disadvantage risk-averse students and bias results due to test-taking strategies. To enhance reliability and more accurately assess ability, Dutch medical schools recently transitioned to a computer adaptive-PT (CA-PT) based on Item Response Theory, which adjusts question difficulty dynamically, excluding the question mark option. This provided a unique opportunity to evaluate the impact of the question mark option in a large cohort. We specifically explored the relationship between question mark use in conventional PT and performance on CA-PT.

Methods: Retrospective data from medical students across seven faculties who took both PT formats were analyzed. Z-scores for total score and question mark score (number of unanswered questions) in the conventional PT, and theta score for the CA-PT were assessed. A linear model assessed the effect of the question mark score on theta, corrected for the conventional PT-score. Cluster analysis explored student subgroups per year.

Results: Students with similar conventional PT scores who left more questions unanswered on the conventional PT generally performed better on CA-PT. This effect diminished as students advance through their studies. Cluster analysis revealed a variable effect between different students, most pronounced in year 4, and a reversed effect in year 5.

Discussion: Question mark option use significantly impacts student performance on PT, with a remarkable variability among students. This variability suggests that formula scoring captures more than knowledge alone, highlighting the need to align scoring methods with intended assessment goals.

Introduction

In classical test theory (CTT), number-right scoring and formula scoring are the two primary scoring methods for calculating test scores on multiple-choice question (MCQ) assessments [1]. Number-right scoring awards points for correct answers without penalizing incorrect ones, while formula scoring deducts points for incorrect answers to discourage guessing. Additionally, formula scoring includes a question mark option, allowing students to acknowledge gaps in their knowledge without penalty [2]. This method has been widely used in medical progress tests (PT) across the Netherlands, Germany, Canada, and at the United Kingdom [3].

The rationale for adopting formula scoring in progress testing is twofold: to encourage students to reflect on how certain they are of their knowledge and to provide an opportunity to indicate when they are unsure, thereby reducing the influence of guessing, particularly for early-stage students who are not yet expected to perform at an end-of-curriculum level [1-5]. However, the question mark option may disadvantage risk-averse students, and introduce biases related to metacognitive skills, self-efficacy beliefs, and test-taking strategies [1, 4, 6-8]. For example, risk-averse students may score lower than equally knowledgeable peers who take more risk, as they tend to use the question mark option more frequently [4, 9-11]. Beyond individual test-taking behaviour, question mark option usage also depends on item-related factors. When clearly incorrect answers can be identified, students have a better chance of guessing correctly from the remaining options and scoring higher, rather than selecting the question mark, which yields 0 points. Moreover, gender differences in guessing behaviour further challenge the construct validity of test scores under formula scorning [4, 10, 12].

To improve the efficiency and reliability of the PT, the Dutch medical schools transitioned form a linear-fixed PT with formula scoring to a computer adaptive progress test (CA-PT) [13]. The CA-PT is based on the Item Response Theory (IRT), which, unlike the CTT, does not assume that all items contribute equally to a student's score. Instead, it uses item difficulty parameters of a set of calibrated questions to measure student's ability ('*theta*'), without the need of a question mark option [14, 15]. In the CA-PT, the difficulty level of the selected questions is adapted based on previous answers, providing a more accurate evaluation of student knowledge [14].

This transition provides a valuable opportunity to evaluate the impact of the question mark option in formula scoring on student performance within a large, summative, and longitudinal cohort of medical students at different educational stages. As students completed both PTs with a question mark option (conventional PT) and without it (CA-PT), this setting allows for a unique examination of how the question mark option influences students' performance. These insights can guide the selection of formula scoring as scoring method. Therefore, the primary aim of this study was to explore the relationship between the question mark option in the conventional PT and student performance on the CA-PT. To establish the validity of this comparison, we first assessed the correlation between these formats over time. We hypothesized that the dismissal of the question mark option would have the greatest impact on the performance of junior students, as they tend to use the question mark option more frequently and have not yet learned how to use this option effectively. They may also be less convinced about their knowledge and therefore answer less questions. This combination makes them more inclined to answer too few

questions, potentially boosting their performance when the question mark option is removed.

Methods

Setting

In the Netherlands, eight universities offer medical education, each with a comparable curriculum structure comprising six years of undergraduate medical education. This curriculum is divided into a three-year (preclinical) Bachelor's program followed by a three-year (clinical) Master's program. The framework for undergraduate medical education defines the joint learning outcomes for both the preclinical and clinical phases, and is applicable to all medical students [16]. The preclinical phase primarily focuses on establishing a theoretical foundation and providing some essential basic skills, while the clinical phase is characterized by clinical rotations. The Dutch interuniversity medical PT is a longitudinal, comprehensive test that evaluates the development of students' functional medical knowledge throughout the entire curriculum, benchmarking against peers at the same stage of study. There are four test administrations (i.e., test moments) each of six academic years (September, December, February, and May) in which medical students of all eight Dutch medical schools participate. This leads to 24 test moments for each student throughout the curriculum. As students progress through their academic years, the passing scores of the PT increase correspondingly. At the end of each academic year, the results from the four progress tests are combined into a summative decision (fail, pass, or good) [17]. To ensure content validity, the PT questions are administered according to a blueprint that prescribes the distribution of questions across relevant medical disciplines (*Appendix 1 – Supplemental Table 1*).

Conventional progress test and formula scoring

The conventional PT was a linear-fixed test format, based on CTT [18]. It comprised 200 multiple-choice questions (MCQs), which featured a question mark option. Choosing this option resulted in a neutral score of zero points. Conversely, selecting an incorrect answer incurs a penalty, leading to a negative score. This formula scoring method was employed to calculate the total score, expressed as a percentage of the maximum score [2, 13].

Computer adaptive progress test (CA-PT)

The CA-PT was introduced across all Dutch medical schools following our cross-over study in May 2022, in which we found a strong correlation between conventional PT and CA-PT test performance [13]. The CA-PT consists of 135 MCQs, of which 120 are calibrated questions, and 15 are non-adaptive pretest questions. The algorithm of the CA-PT selects questions from an item bank. At time of data collection, this item bank comprised 5,613 calibrated questions. Pretest questions are included for calibration purposes to be used in future tests. The score of the CA-PT is represented as an estimated ability level (theta score), calculated based on the answers to the 120 calibrated questions in conjunction with their difficulties [19].

Participants

We used retrospective data from medical students at seven Dutch medical schools who participated in both the conventional PT and the CA-PT. Data from one medical school were excluded because this institution joined the PT after implementation of the CA-PT, resulting in a lack of conventional PT data.

Study design and data collection

We used data from four conventional PTs administered between 2021 and 2022 (September 2021, December 2021, February 2022, and May 2022), and from five CA-PTs administered between 2022 and 2023 (May 2022, September 2022, December 2022, February 2023, and May 2023) (*Figure 1*). The PT results from May 2022 origin from our cross-over study [13], in which 1,432 students from three Dutch medical schools participated in both a conventional PT and CA-PT in the same time frame. For each PT, we extracted the test moment, medical school, and student ID. For the conventional PTs, we used the total PT score, and the "*question mark score*" (i.e., the number of questions left open by the student). For the CA-PTs, we used the theta score. Due to COVID-19 restrictions in 2021-2022, some conventional PT-sessions were conducted online for part of the students and were non-proctored (*n*=8,021/37,412 PTs in 2021-2022). These "*formative*" PT-sessions, which did not impact study credits due to the lack of supervision on the use of study materials, were excluded from our analyses as previous findings indicated that their purely formative nature could affect test-taking motivation and student performance [20].

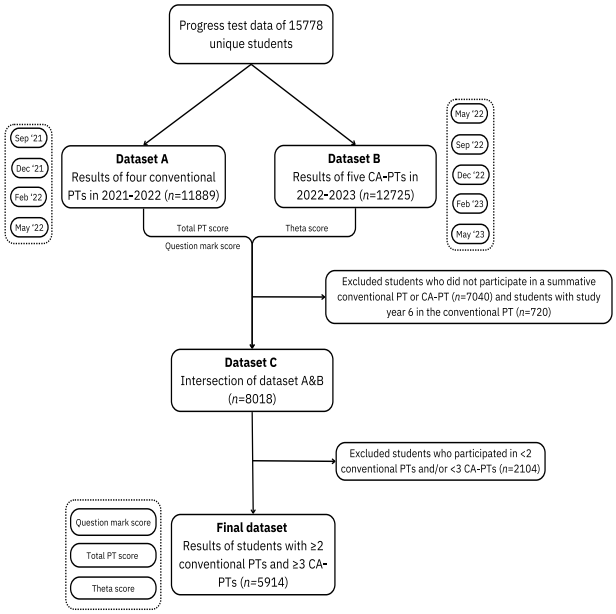


Figure 1. Flowchart of the data collection

We linked the conventional PT data with the CA-PT data by student ID and selected data from students who participated in at least two conventional PTs and three CA-PTs (*n*=5,914). We used a lower threshold for including the conventional PT (2 *versus* 3), because we had to exclude the "*formative*" sessions, which would otherwise have limited the number of students eligible for this study. Each student was assigned a year group (1-6, representing the cohort) based on their earliest test moment in 2021-2022. Students in year group 6 were excluded due to the small, non-representative sample attending both the conventional PT and the CA-PT.

Data analysis

We calculated average z-scores for both the total score and question mark score on the conventional PT, relative to all students in the same test moment group for each separate PT session. Specifically, we computed the z-scores for each student's total score and question mark score, then averaged these z-scores across all students within the same test moment group. This z-score corrects for differences in test difficulty and accounts for individual student growth over time, providing a measure of the student's longitudinal position within the population regarding both total score and the number of unanswered questions. For the CA-PT, we calculated the z-score for the theta-score. Since the CA-PT sessions are comparable in terms of difficulty level, we combined all CA-PT sessions to determine the test moment groups, from which we calculated z-scores. We computed descriptive statistics for the selected student groups, including density plots to visualize the relationships between the different variables.

Correlation between conventional PT and CA-PT results

We computed the Pearson correlation coefficient to assess the convergent validity of the conventional PT and CA-PT by measuring the correlation between the total score on the conventional PT and the theta score on the CA-PT across multiple test sessions longitudinally.

Relationship between question mark option use and CA-PT performance

We investigated the relationship between the question mark score in the conventional PT and the theta score in the CA-PT using regression analysis and model-based clustering. Linear regression models were applied for each year group to correct for the conventional PT score and assess the effect of the question mark score on the theta score. Because we observed signs of underlying structure in the data, we decided to apply model-based clustering to further explore the subgroups in the year groups using the R package MClust, (version 5) [21]. This approach identified clusters of students with distinct patterns in their question mark and theta scores. Optimal model class and cluster numbers were determined using the Bayesian Information Criterion (BIC), and Integrated Complete-data Likelihood (ICL), and Likelihood Ratio Tests (LRT) bootstrapping. Clusters were subsequently included as covariates in linear regression models to evaluate differences in behaviour between clusters within each year group. All statistical analyses were performed in R version 4.1.0 [22].

Ethical approval

We used data from our cross-over study conducted in May 2022, for which ethical approval was granted by the Ethical Review Board of the Netherlands Association for Medical Education (NVMO) under reference NERB/2023.4.6. Participation in this CA-PT (May 2022) was voluntary, with students being informed beforehand and providing signed informed consent prior to its initiation. Retrospective data from other PT sessions were obtained from the PT database, which is maintained for the purposes of monitoring and improving PT administration. A waiver for the use of this retrospective data was granted by the NVMO Ethical Review Board. All data were pseudonymized before analysis.

Results

Descriptives

We included 5,914 students who participated in at least two summative conventional PTs and three CA-PTs for analysis (*Figure 1*). The number of students who participated in each possible combination of conventional PT and CA-PT sessions are shown in *Supplemental Table 2* (*Appendix 2*). In year group 3 we had only 415 students, because their participation in (CA-)PTs was reduced due to disruptions caused by the COVID-19 pandemic (i.e., they experienced a longer waiting period between the pre-clinical and clinical phase during which they did not participate in PTs). There were slight but statistically significant differences in the average z-scores per test moment in the bachelor phase of the selected students and the total student population. The differences ranged from -0.25 to +0.25, but were not systematic (*Appendix 3 – Supplemental Table 3*). *Supplemental Table 4* (*Appendix 4*) presents the mean absolute PT, question mark, and theta scores for each year group, to illustrate students' test-taking behaviour across the year groups.

Correlation between conventional PT and CA-PT results

We observed an overall Pearson correlation of 0.74 between the average z-score on the conventional PT and the CA-PT. The correlation was moderate to strong within each year group: 0.57 (Y1; *n*=1,067), 0.71 (Y2; *n*=1,017), 0.70 (Y3; *n*=415), 0.79 (Y4; *n*=2,615), and 0.80 (Y5; *n*=800).

Relationship between question mark option use and CA-PT performance

All results are presented as average z-scores. For simplicity and readability, however, we will refer to these values as "*scores*" throughout the remainder of this paper. Our linear model revealed a significant interaction between question mark score and theta score across all year groups. As illustrated in *Figure 2*, the question mark score (x-axis) positively affects the theta score (y-axis), after adjusting for the total score on the conventional PT (PT score). This positive effect suggests that students who left more questions unanswered on the conventional PT, indicating greater uncertainty, tended to perform better on the CA-PT for a given PT score. The positive effect was strongest in year group 1, but decreased as students progressed through their studies (Y1: 0.47; Y2: 0.39; Y3: 0.32; Y4: 0.24; Y5: 0.22).

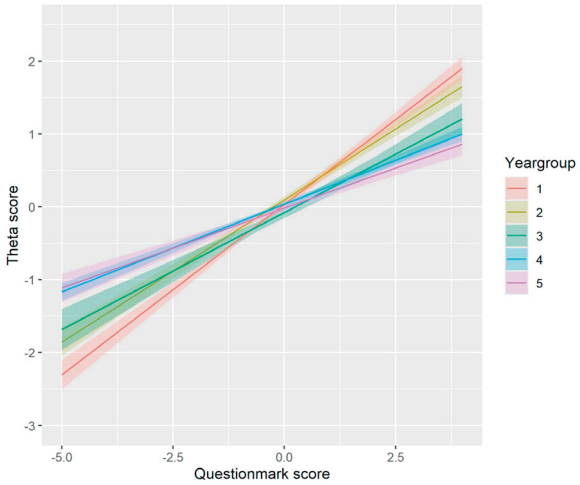


Figure 2. The effect of the question mark score (conventional PT) on the theta score (CA-PT), corrected for the PT score, for each year group. Positive theta scores indicate better-than-average performance, suggesting the individual is performing above the mean level. Negative theta scores suggest below-average performance, indicating the individual is performing below the mean level. The question mark score quantifies how frequently a student uses the question mark option. Higher scores indicating more frequent use while lower or negative scores indicate a preference for direct answers.

To examine the observed underlying structure in our data, we applied model-based clustering with LRT bootstrapping. This analysis revealed the underlying structure of the effects across the year groups. Four clusters were identified in year groups 1, 2 and 5, while six clusters emerged in year group 4. Year group 3 did not exhibit distinct clustering. As our data is three-dimensional, *Figure 3* presents the effect of question mark scores on theta scores within each cluster from three perspectives: **A)** question mark score versus theta score; **B)** question mark score versus PT score; **C)** and theta score versus PT score. Each data point in the graph represents a student. For each student, we assessed three variables: the CA-PT score, question mark score, and PT score, which are displayed in three separate projections to better visualize the cluster formations. Within each cluster, we applied a linear model to examine the effect of question mark score on theta score, while adjusting for the PT score. The colour of each cluster indicates the effect size of this relationship. The shape and position of the clusters represent the type of students in each cluster based on their scores. For example, in cluster 3 of year group 1 (shown in purple), students had high question mark scores with a low variability (as seen in *Figure 3 graph A* and *B*), while their theta scores showed much greater variability (*Figure 3 graph A*). In this cluster, the linear model revealed a strong positive effect (indicated by red datapoints) of question mark option usage on the theta score, after adjusting for PT score. These effects are also shown in *Figure 4* by different line colours for each cluster.

Unlike *Figure 2*, which shows a consistent positive effect across all year groups, *Figure 3* and *4* reveal a more nuanced pattern; positive effects of the question mark score on the theta score for the clusters in first and second year group, but predominantly negative effects for the clusters in the fifth year group. This suggests that for less experienced students, greater reliance on the question mark option – leaving more questions unanswered – is associated with better performance on the CA-PT. Conversely, in more

experienced students, higher use of the question mark option correlates with poorer CA-PT performance. Year group 4 showed the greatest variability in cluster effects (*Figure 3, 4*). In the following paragraphs we will describe in more detail the clusters that stood out. Year group 3 was excluded from this analysis due to insufficient data for clustering. *Supplemental Table 5* (*Appendix 5*) shows the mean scores (PT, CA-PT, and question mark), and effect sizes for each cluster.

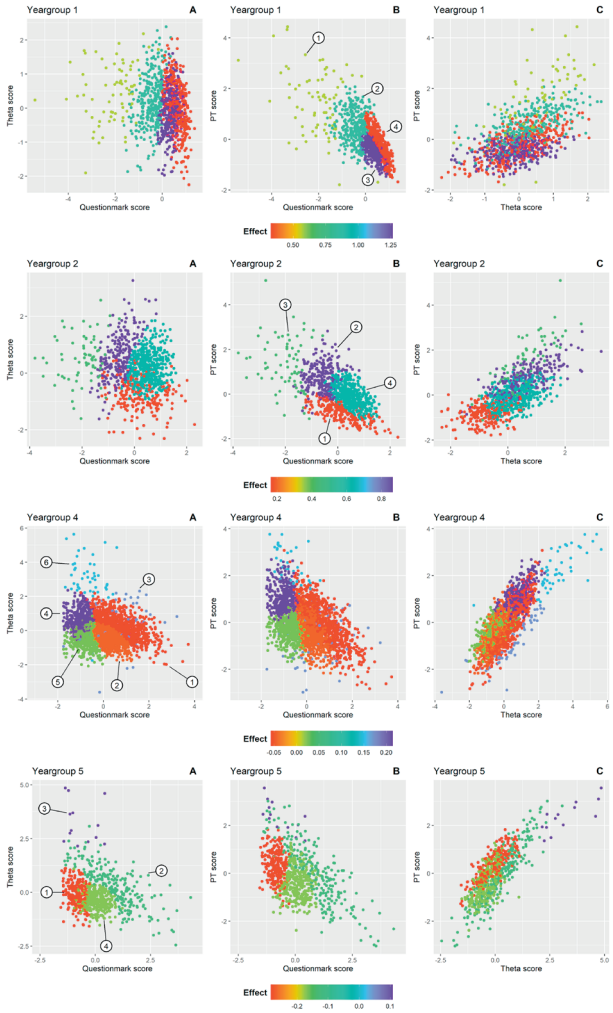


Figure 3. The effect of the question mark score (conventional PT) on the theta score (CA-PT), corrected for the PT score for the clusters in within each your group shown from three perspectives: **A)** question mark score versus theta score **B)** question mark score versus PT score **C)** theta score versus PT score. The cluster numbers of each year group are shown in the graph where the clusters are most distinctly separated. Cluster colors represent the effect size of the effect of question mark score on the theta score, adjusted for PT score. Red indicates the most positive effect and purple the most negative effect. The color code is consistent across all graphs within each year group. The color scale differs in range for each year group.

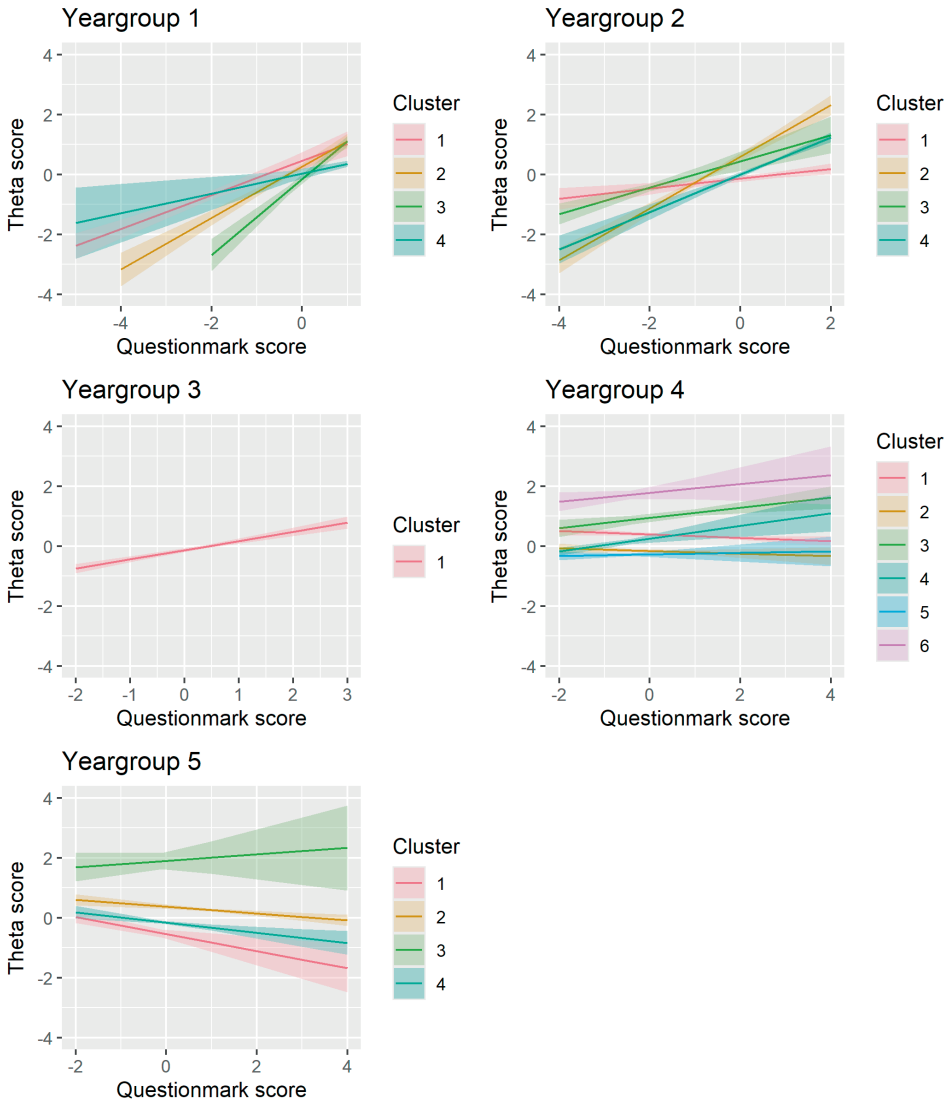


Figure 4. The effect of the question mark score (conventional PT) on the theta score (CA-PT), corrected for the PT score, within each cluster across the five year groups. Each cluster is indicated with a different colour. Optimal clustering solutions were identified using the Bayesian Information Criterion (BIC) and Integrated Complete-data Likelihood (ICL). BIC evaluates the model fit while penalizing complexity, ensuring an appropriate balance between accuracy and simplicity. ICL complements this by emphasizing well-separated and distinct clusters, reducing the risk of overfitting. Together, these criteria provided a framework to determine the number of clusters in each year group.

Year group 1

In year group 1, students in ***cluster 3*** (*n*=324) and ***cluster 4*** (*n*=360) used the question mark option more than average in the conventional PT, with ***cluster 4*** students using it more often (mean [SD]: ***cluster 3***: 0.34 [0.26]; ***cluster 4***: 0.63 [0.29]). These clusters were difficult to separate clearly based on their question mark score and theta score, as shown by the overlap between clusters in *Figure 3A*. Their scores on the CA-PT were similar (***cluster 3***: -0.13 [0.63]; ***cluster 4***: 0.03 [0.74]). However, their scores on the conventional PT differed with cluster 4 students performing better (***cluster 3***: -0.45 [0.36]; ***cluster 4***: -0.21 [0.58]). The key difference was in how effectively they used the question mark option: in ***cluster 3***, abandoning the question mark option strongly boosted CA-PT performance as illustrated by the steepest line (green) in *Figure 4* (i.e., strong positive effect). In ***cluster 4***, the effect was much weaker (purple line in *Figure 4*). This suggests that students in ***cluster 4*** were better at using the question mark option to improve their performance on the conventional PT compared to those in ***cluster 3***.

Year group 2

Students in ***cluster 2*** (*n*=243) used the question mark option below average (-0.58 [-0.44]), while achieving the highest mean CA-PT score (0.61 [-0.76]). Despite their limited use of the question mark option in relation to their peers, their CA-PT performance was boosted following the removal of the question mark option, as illustrated by the steep line in *Figure 4* (i.e., strong positive effect). This suggests that they did not use the question mark option effectively, and the reliance on direct answers in the CA-PT resulted in a better performance. Similarly, students from ***cluster 4*** (*n*=481) who exhibited the greatest uncertainty and the highest question mark usage (0.51 [0.41]) gained from the removal of the question mark option. These students also show a great improvement from a below-average PT score (-0.12 [0.47]) to an above-average CA-PT score (0.20 [0.53]). This suggests that while these students initially depended heavily on the question mark option, its removal encouraged more decisive responses, enhancing their CA-PT performance.

Year group 4

In this year group, students in ***cluster 4*** (*n*=411) answered most questions directly on the PT, reflected by the lowest question mark score (-0.99 [0.32]). Their strategic use of the question mark option resulted in the highest mean conventional PT score (1.02 [0.52]). The dismissal of the question mark option even further boosted their CA-PT performance (0.68 [0.55]), reflected by the positive effect (*Figure 4*). In contrast, students in ***cluster 5*** (*n*=518) experienced neither a benefit nor a disadvantage from the removal of the question mark option, as reflected by a near-neutral effect (horizontal slope of the blue line in *Figure 4*). This suggests that these students were well aware of their knowledge gaps and used the question mark option effectively.

Year group 5

Students in ***cluster 1*** (*n*=236) answer most questions directly on the conventional PT, reflected by the lowest question mark score (-0.85 [0.29]). They scored above average on the conventional PT (0.33 [0.61]), but their performance on the CA-PT was below average (-0.10 [0.57]). This pattern, together with the strong negative effect, suggests that these students used the question mark option strategically to improve their score on the conventional PT. However, without the question mark option on the CA-PT, their scores reveal a lower knowledge level that resulted in below-average CA-PT performance.

Discussion

We found a strong correlation of average z-scores between the two PT formats over time, supporting their convergent validity and strengthening the justification for the switch to CA-PT [13]. This result allows score comparisons across the different PT formats, suggesting that observed score differences are primarily due to question mark usage. The overall effect of the question mark score on the theta score, adjusted for the PT score, was consistently positive across all year groups but diminished with student progression. While the general trend was positive, our cluster analysis exposed varying student behaviours within each year. Notably, year group 4 showed considerable variation in student behaviour, and in year group 5 the effect reversed, becoming predominantly negative.

The effect of question mark use on the theta score shifts notably over the curriculum, shifting from a strong positive effect in year group 1 to a predominantly negative effect in year group 5. The negative effects were evident in the clusters, but not in the overall effect of the entire year 5 cohort, possibly due to the Simpson's paradox, where aggregated data obscure subgroup trends [23]. This shift may indicate the development of students' test-taking strategies, and higher pass thresholds of the PT in advanced years.

Early in the curriculum, frequent question mark option use on the conventional PT correlated with higher CA-PT scores, possibly reflecting differences in metacognitive skills and self-efficacy beliefs, as noted in previous studies [1, 4]. Students who answered fewer questions on the conventional PT, but performed well, may be more risk-averse and benefitted from the mandatory answering format of the CA-PT. Conversely, students who took more risk on the conventional PT and used the question mark option less frequently but effectively, improved their conventional PT score. However, without the question mark option on the CA-PT, their score was lower. Despite differences in knowledge level between these student groups, the conventional PT yields similar scores due to differences in risk tendency and test-taking strategies, but leads to different results on the CA-PT.

In higher year groups, students answering fewer questions on the conventional PT fared worse on the CA-PT, likely due to limited knowledge masked by strategically question mark use. In the CA-PT, where all questions must be answered, their lower knowledge levels become more apparent. Cecilio-Fernandes *et al.* [24] found that the accuracy of students' judgement of knowledge decreases over time, with students providing both more correct and incorrect answers as they progress through the medical program. Our results partly suggest the same, as the students with the lowest question mark score within year group 5 have a lower mean score on the conventional PT compared to clusters in earlier years with low mean question mark scores (e.g., cluster 1 in year group 5 *vs.* cluster 2 in year group 2). This might reflect a change in answering strategies or an increase in risk-taking behaviour. However, year group 4 showed no consistent relationship between question mark use and CA-PT performance, suggesting a heterogenous group of students with varying behaviours. While the overall effect remained predominantly positive, we also observed clusters with a near-neutral effect. This variation may reflect the development of effective test-taking strategies (e.g., effective question mark option use), self-assessment, and different prior trajectories in this year group [24]. Some students in year 4 transitioned directly from year 3, while others completed a research internship or pursued other activities before starting clinical rotations, causing heterogeneity among the students in this year group. Finally, students who tended to answer more

questions on the conventional PT generally achieved higher scores on both PT formats. This effect was most pronounced in the relatively small, but clearly distinguished, clusters of best-performing students, whose scores were less affected by question mark usage (e.g., cluster 3 in year group 2). Overall, the wide diversity in student behaviour observed across and within year groups in our study suggests that formula scoring assesses constructs beyond knowledge level, including metacognitive awareness and risk-tendencies.

Strengths and limitations

The strengths of this study include its multi-center design, the large cohort of medical students at different stages in the medical curriculum, and the use of summative PT results, minimizing selection bias regarding student participation. The longitudinal design provided a nuanced understanding of formula scoring effects on student performance. This study also faced limitations, including potential selection bias favoring higher-performing students as the selected cohort had slight but statistically significant differences in z-scores compared to the total population. Additionally, lack of access to student characteristics hindered a more in depth-analysis of the underlying mechanisms or traits driving the observed student behaviour. Consequently, our interpretation of the underlying behaviour explaining the observed cluster scores and effects are speculative and based on earlier research. The differences in test formats of the conventional PT and the CA-PT (flexible navigation through the questions vs. direct answering format) may have influenced student strategies and performance, complicating direct comparisons. Cluster analysis sensitivity to input data, outliers, and nondeterminism, may have influenced the clusters identified, particularly where overlapping clusters with similar scores exhibited different effects. While this may have affected individual cluster assignments, we anticipate that it did not significantly impact the overall observed group-level patterns. However, some clusters exhibited large score variances, making it difficult to draw definitive conclusions.

Implications and future research

Our results support and expand on prior research that formula scoring affects construct validity of test scores [4, 10]. The high variability in question mark use and its impact suggest that formula scoring introduces bias, potentially distorting knowledge measurement. Given this, using formula scoring solely to administer the same test to students with diverse proficiency levels, as in the PT, may be unjustified. If the goal is to assess metacognitive knowledge [25], formula scoring could be appropriate, but alternative self-assessment methods like certainty-based marking (CBM) [26] might better account for students' confidence in their answers and improve their self-reflection [27, 28]. Future research could explore the underlying traits or mechanisms behind the behavioural differences and outcomes across student clusters through qualitative methods such as interviews.

Conclusion

Our study demonstrates that question mark option use in formula scoring significantly influences student performance on the PT, with the effect varying across different stages of the curriculum. The great variability suggests that formula scoring measures not only knowledge, but also other student constructs, potentially introducing biases. Careful consideration of scoring methods aligned with the assessment goals is essential to ensure valid and reliable test outcomes.

References

1. 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.
2. Lord FM. Formula scoring and number-right scoring. Journal of Educational Measurement. 1975;12(1):7-11.
3. 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.
4. 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.
5. Rowley G, Traub R. Formula scoring, number-right scoring, and test-taking strategy. Journal of Educational Measurement. 1977;14(1).
6. 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.
7. Rowley GL, Traub RE. Formula Scoring, Number-Right Scoring, and Test-Taking Strategy. Journal of Educational Measurement. 1977;14(1):15-22.
8. 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).
9. Lord FM, Lord FM. Formula Scoring and Validity. Educational and Psychological Measurement. 1963-12-01;23(4).
10. 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.
11. Kampmeyer D, Matthes J, Herzig S. Lucky guess or knowledge: a cross-sectional study using the Bland and Altman analysis to compare confidence-based testing of pharmacological knowledge in 3rd and 5th year medical students. Advances in Health Sciences Education. 2015;20(2):431-40.
12. Budescu D, Bar-Hillel M. To Guess or Not to Guess: A Decision-Theoretic View of Formula Scoring. Journal of Educational Measurement. 1993/12/01;30(4).
13. van Wijk EV, Donkers J, de Laat PCJ, Meiboom AA, Jacobs B, Ravesloot JH, et al. Computer Adaptive vs. Non-adaptive Medical Progress Testing: Feasibility, Test Performance, and Student Experiences. Perspectives on Medical Education. 2024;13(1).
14. Chang H-H. Psychometrics behind Computerized Adaptive Testing. Psychometrika. 2015;80:1-20.
15. Downing SM. Item response theory: applications of modern test theory in medical education. Medical Education. 2003;37(8):739-45.
16. Framework for Undergraduate Medical Education 2021 [updated 2021-08-20; cited July 2023]. Available from: https://www.nfu.nl/en/themes/professional-future/medicine-programmes/framework-undergraduate-medical-education.
17. 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.
18. Traub RE. Classical Test Theory in Historical Perspective. Educational Measurement: Issues and Practice. 2005;16(4):8-14.
19. Warm TA. Weighted likelihood estimation of ability in item response theory. Psychometrika. 1989;54(3):427-50.
20. van Wijk EV, van Blankenstein FM, Donkers J, Janse RJ, Bustraan J, Adelmeijer LGM, et al. Does 'summative' count? The influence of the awarding of study credits on feedback use and test-taking motivation in medical progress testing. Advances in Health Sciences Education 2024;29:1665-1688.
21. Fraley C, Raftery AE. Model-Based Clustering, Discriminant Analysis, and Density Estimation. Journal of the American Statistical Association. 2002;97(458).
22. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2021 [cited 2024]. Available from: https://www.R-project.org/.
23. Simpson E. The interpretation of Interaction in Contingency Tables. Journal of the Royal Statistical Society Series B (Methodological). 1951;12(2):3.
24. Cecilio-Fernandes D, Kerdijk W, Jaarsma ADC, Tio RA. Development of cognitive processing and judgments of knowledge in medical students: Analysis of progress test results. Medical teacher. 2016;38(11).
25. Krathwohl DR. A Revision of Bloom's Taxonomy: An Overview. Theory Into Practice. 2002;41(4).
26. Gardner-Medwin AR. Confidence assessment in the teaching of basic science. Research in Learning Technology. 1995;3(1).
27. Cash B, Mitchner NA, Ravyn D. Confidence-Based Learning CME: Overcoming Barriers in Irritable Bowel Syndrome With Constipation. Journal of Continuing Education in the Health Professions. 2011;31(3).
28. Luetsch K, Burrows J. Certainty rating in pre-and post-tests of study modules in an online clinical pharmacy course - A pilot study to evaluate teaching and learning. BMC Medical Education. 2016;16(1).

Appendix

Appendix 1 – Supplemental Table 1

Supplemental Table 1A. Blueprint used in the conventional progress test.

Discipline	Number of questions
Anatomy	13
Biochemistry/Genetics/Histology/Molecular Cell Biology	18
Surgery	17
Dermatology/Ear, Nose, Throat/Ophthalmology	14
Epidemiology/Statistics	8
Pharmacology	9
Physiology	11
Geriatrics	8
Gynecology/Obstetrics	7
General Practice	20
Internal Medicine	26
Pediatrics	12
Metamedics	5
Neurology	7
Pathology/Immunology/Microbiology	10
Psychiatry/Psychology	12
Social Medicine	3
Total	200

Supplemental Table 1B. Blueprint used in the computer adaptive progress test.

Discipline	Number of questions
Anatomy	7;8
Biochemistry/Genetics/Histology/Molecular Cell Biology	10;11
Surgery	10;11
Dermatology/Ear, Nose, Throat/Ophthalmology	8;9
Epidemiology/Statistics	4;5
Pharmacology	5;6
Physiology	6-7
Geriatrics	4-5
Gynecology/Obstetrics	4-5
General Practice	12
Internal Medicine	15-16
Pediatrics	7-8
Metamedics	3
Neurology	4-5
Pathology/Immunology/Microbiology	6
Psychiatry/Psychology	7-8
Social Medicine	1-2
Total	120

Appendix 2 – Supplemental Table 2

Supplemental Table 2. The number of students who participated in various combinations of the conventional PT and CA-PT.

Conventional progress test	Computer adaptive progress test				
	1	2	3	4	5
1	48	141	216	374	45
2	145	176	***389***	***950***	***190***
3	197	338	***827***	***2115***	***365***
4	240	184	***253***	***655***	***170***

The students included in our analyses are in italic bold.

Appendix 3 – Supplemental Table 3

Supplemental Table 3. The differences in average z-scores for the theta-score (CA-PT), PT-score (conventional PT), and question mark score between our study population and the entire student population expressed as p-values of independent t-tests.

Year	Test moment	Theta score	PT score	Question mark option score
1	1	*NA*[a]	***<0.001***	***<0.001***
	2	*NA*	0.51	0.83
	3	*NA*	0.16	***<0.001***
	4	*NA*	0.06	0.27
2	5	***0.02***	0.87	***<0.001***
	6	0.64	0.46	0.96
	7	***0.01***	0.32	0.03
	8	0.06	0.37	0.73
3	9	***<0.001***	***0.01***	0.94
	10	0.83	0.28	***0.005***
	11	***<0.001***	***<0.001***	0.04
	12	0.79	0.14	0.31
4	13	0.06	0.66	0.95
	14	0.92	0.30	0.44
	15	0.24	0.39	0.37
	16	0.82	0.85	0.65
5	17	0.28	0.49	0.70
	18	0.47	0.95	0.90
	19	0.73	0.64	0.53
	20	0.74	0.72	0.46
6	21	0.48	*NA*	*NA*
	22	0.47	*NA*	*NA*
	23	0.48	*NA*	*NA*
	24	0.26	*NA*	*NA*

[a]Not applicable due to small samples. P-values <0.05 in italic bold.

Appendix 4 – Supplemental Table 4

Supplemental Table 4. Mean raw scores on the CA-PT (theta score), conventional PT (PT score), and question mark option

Year group	Theta scorea	PT score	Question mark score
1 (*n*=1067)	-0.16 (0.20)	10.41 (3.90)	76.03 (11.05)
2 (*n*=1017)	0.04 (0.20)	20.39 (5.85)	54.45 (12.56)
3 (*n*=415)	0.11 (0.20)	30.04 (8.10)	33.59 (13.69)
4 (*n*=2615)	0.29 (0.19)	38.95 (8.87)	20.80 (11.34)
5 (*n*=800)	0.33 (0.17)	44.26 (8.61)	14.27 (9.22)

[a]All scores are expressed as mean (standard deviation). PT-score and question-mark score are on a scale of 0-100. Theta scores (CA-PT) are on a logit scale.

Appendix 5 – Supplemental Table 5

Supplemental Table 5. Mean z-scores on the CA-PT (theta score), conventional PT (PT score), question mark option (question mark score) and the effect size for each cluster across the five year groups.

Year group	Cluster	Theta score[a]	PT score	Question mark score	Effect
1	1 (*n*=74)	0.47 (0.82)	1.48 (1.26)	-2.06 (0.94)	0.57
	2 (*n*=309)	0.24 (0.75)	0.50 (0.64)	-0.46 (0.38)	0.86
	3 (*n*=324)	-0.13 (0.63)	-0.45 (0.36)	0.34 (0.26)	1.26
	4 (*n*=360)	0.03 (0.74)	-0.21 (0.58)	0.63 (0.29)	0.33
2	1 (*n*=221)	-0.76 (0.54)	-0.86 (0.38)	0.20 (0.70)	0.17
	2 (*n*=243)	0.61 (0.76)	0.74 (0.61)	-0.58 (0.44)	0.86
	3 (*n*=72)	0.53 (0.80)	1.33 (1.16)	-1.95 (0.67)	0.44
	4 (*n*=481)	0.20 (0.53)	-0.12 (0.47)	0.51 (0.41)	0.62
3	1 (*n*=415)	-0.10 (0.87)	-0.06 (0.89)	0.13 (0.90)	0.31
4	1 (*n*=708)	0.33 (0.73)	-0.02 (0.94)	0.86 (0.76)	-0.06
	2 (*n*=875)	-0.39 (0.45)	-0.33 (0.63)	0.19 (0.37)	-0.04
	3 (*n*=57)	0.49 (1.17)	-0.85 (0.91)	0.54 (1.00)	0.17
	4 (*n*=411)	0.68 (0.55)	1.02 (0.52)	-0.99 (0.32)	0.21
	5 (*n*=518)	-0.48 (0.52)	-0.28 (0.46)	-0.73 (0.33)	0.03
	6 (*n*=46)	3.11 (1.21)	2.25 (0.80)	-0.66 (0.57)	0.15
5	1 (*n*=236)	-0.10 (0.57)	0.33 (0.61)	-0.85 (0.29)	-0.28
	2 (*n*=231)	0.30 (0.93)	0.01 (1.23)	0.89 (1.06)	-0.11
	3 (*n*=13)	3.22 (0.98)	2.36 (0.63)	-0.65 (0.69)	0.11
	4 (*n*=320)	-0.39 (0.40)	-0.42 (0.62)	0.05 (0.41)	-0.17

[a]All scores are expressed as mean (standard deviation). PT: (conventional) progress test.

Part III: Feedback

Chapter 8

Does ‘summative’ count? The influence of the awarding of study credits on feedback use and test-taking motivation in medical progress testing

Elise V. van Wijk
Floris M. van Blankenstein
Jeroen Donkers
Roemer J. Janse
Jacqueline Bustraan
Liesbeth G.M. Adelmeijer
Eline A. Dubois
Friedo W. Dekker
Alexandra M.J. Langers

Advances in Health Science Education. 2024;29(5):1665-1688.
DOI: 10.1007/s10559-10324-4.

Abstract

Despite the increasing implementation of formative assessment in medical education, its' effect on learning behaviour remains questionable. This effect may depend on how students value formative, and summative assessments differently. Informed by Expectancy Value Theory, we compared test preparation, feedback use, and test-taking motivation of medical students who either took a purely formative progress test (*formative* PT-group) or a progress test that yielded study credits (*summative* PT-group). In a mixed-methods study design, we triangulated quantitative questionnaire data (n = 264), logging data of an online PT feedback system (n = 618), and qualitative interview data (n = 21) to compare feedback use, and test-taking motivation between the *formative* PT-group (n = 316), and the *summative* PT-group (n = 302). Self-reported, and actual feedback consultation was higher in the *summative* PT-group. Test preparation, and active feedback use were relatively low and similar in both groups. Both quantitative, and qualitative results showed that the motivation to prepare and consult feedback relates to how students value the assessment. In the interview data, a link could be made with goal orientation theory, as performance-oriented students perceived the *formative* PT as not important due to the lack of study credits. This led to low test-taking effort, and feedback consultation after the *formative* PT. In contrast, learning-oriented students valued the formative PT, and used it for self-study or self-assessment to gain feedback. Our results indicate that most students are less motivated to put effort in the test, and use feedback when there are no direct consequences. A supportive assessment environment that emphasizes recognition of the value of formative testing is required to motivate students to use feedback for learning.

Introduction

The notion that *'assessment drives learning'* is widely acknowledged [1, 2]. The way learning is driven may therefore differ with the focus of the assessment. Within medical education, the focus is mainly on assessment of learning [3]. These summative assessments evaluate students' learning by measuring their performance, often reported as a summative grade. In contrast to assessment *of* learning, assessment *for* learning uses formative assessments to enhance in-depth learning, and self-regulated learning (SRL) by providing ongoing feedback [4-13]. More specifically, formative feedback provides opportunities for reflection, identifying learning gaps, and adjusting learning, which are important aspects of SRL [5, 14, 15] In this way, feedback can also stimulate the use of learning strategies that enhance future learning performance [16]. With the growing consensus that assessment should promote learning, and in light of these positive learning effects, there is a shift in assessment *of* learning towards assessments *for* learning in medical education [3, 11]. However, to facilitate this shift, further elucidation of the complex relationship between assessment, learning, and the driving factors behind students' learning is needed.

One of the factors found to drive students' motivation to learn is increasing the weight of summative assessments [17]. Motivation to learn for an assessment also affects test-taking motivation: students' readiness to invest effort in a test [18-21]. Considering the lack of direct consequences of formative test results, students might be less motivated to put their best effort in these tests. This can be explained by the Expectancy-Value Theory (EVT), a conceptual framework frequently used in the context of test-taking motivation. This theory assumes that motivation for a task depends on expectancies of success, and perceived value given to the task [22]. Specifically, motivation for a task increases when people expect to be successful and when they find the task valuable for themselves. Test-taking effort is the main element of test-taking motivation, which, according to the EVT, is thus the direct outcome of expectancy and value. Most studies that investigated EVT in the context of test-taking motivation focused on the value component of EVT. Overall, these studies report positive relationships between value and test performance, and also between test effort and test performance [19, 23].

Another way to look at *'assessment drives learning'* is through the lens of the goal orientation theory. This theory states that the individual goal orientation affects motivation, which in turn guides behavioural responses [24]. Goal orientation can either rely on learning (mastery- or learning-oriented goals) or performance (performance-oriented goals). Learning-oriented students might take a different approach in making a test, and using its feedback than performance-oriented students, but so far the influence of goal orientation in different assessment conditions has not been investigated.

One way to investigate the differences between different assessment conditions is by using the medical progress test (PT), which is a frequently used assessment method in medical education. The PT is a longitudinal, comprehensive, and curriculum-independent test administered repeatedly to assess students' knowledge progress and provide feedback [25, 26]. The PT combines longitudinal testing with feedback, serving an important formative function, but in many educational contexts the results of PTs are also used for a summative pass/fail decision followed by the rewarding of study credits. As the PT covers the entire medical curriculum, it discourages test-directed studying, and encourages self-directed learning by using the feedback of the previous PT [26].

Implementing frequent PTs with a summative component, and the integrating purely formative PTs (no study credits involved) in a curriculum with other formative assessments has shown a positive impact on students' test-effort, perceived learning value, and feedback use [27-31]. However, some studies have not found the expected beneficial impact of feedback in purely formative PTs on learning [26, 32-35]. Different educational conditions affect the test-taking effort, and the perceived value of purely formative PTs [30]. These PTs have no direct consequences (i.e. no *'stakes'*) for study progress, which may lead to a lower perceived value, which in turn may result in less test-taking motivation and effort put in these tests [30, 36]. Besides an impact on test performance, this might also affect their use of feedback.

In summary, while assessment should promote learning (i.e. assessment *for* learning), the actual effect of formative assessments on learning is unclear. More specifically, it remains unclear how formative versus summative assessment affects students' feedback use, and test-taking motivation. The PT provides a unique opportunity to study this distinction, especially when we can compare a purely formative PT with a PT that also has a summative component. Understanding how students adapt their learning behaviour to formative versus summative assessment may help teachers optimize both functions of assessment, as it enables them to react to the student's behaviour in order to promote their learning process, and foster lifelong learning. Therefore, we aimed to investigate the effect of a PT with a summative component (*summative* PT), and a purely formative PT (*formative* PT) on medical students' (1) test preparation, (2) factors that influence test taking motivation, and the use of feedback, and (3) self-reported, and actual feedback use after the test.

Methods

Study design

We used a convergent mixed-methods approach with a subtle realism paradigm, involving a questionnaire, online Progress test Feedback system (ProF) logging data, and semi-structured interviews. The subtle realism paradigm combines a realist ontology (an objective reality independent of our perceptions) with a constructivist epistemology (our understanding of reality depends on our perspectives) [37, 38]. This paradigm aims at representing reality rather than attaining *"the truth"*, by triangulating different data sources, perspectives, and theories. We chose this approach as this best aligns with our research design, which attempts to represent, and deepen our understanding of reality ('feedback use in the context of different assessment conditions') by the triangulating different data sources, and theories. This paradigm allows us to integrate different perspectives while remaining flexible in interpreting our qualitative data. All data types were analysed separately and converged in a final interpretation phase, where we compared the results of the quantitative and qualitative data, and assessed whether the data confirmed or disconfirmed each other. Our qualitative results, using the existing theoretical frameworks of EVT and goal-orientation theory, helped us understand, and explain the observed and self-reported quantitative feedback behaviour.

Setting

The study was conducted at Leiden University Medical Center (LUMC) in the Netherlands. The medical curriculum in the Netherlands includes a three year preclinical Bachelor program and three year clinical Master program. The Bachelor program comprises several theoretical courses, assessed by written

summative assessments at the end of each course, and rewarded with study credits. Most courses also offer a formative assessment for practice, which is not mandatory for students to take. In the Master program, students undergo clinical rotations, assessed by a pass or fail decision based on supervisor feedback. Throughout their six years of medical school, all medical students take four PTs per year, resulting in a total of 24 test moments [25]. The PTs are taken in September (PT 1), December (PT 2), February (PT 3), and May (PT 4). The PT is a comprehensive written test of 200 multiple choice questions (MCQs), covering all relevant medical disciplines, and stratified in categories [26]. The MCQs include a question mark option that yields no points, and points are deducted for incorrect answers [39]. All participating students take an identical PT in an exam hall with proctoring. The final score on the PT is expressed as a percentage of the maximum attainable score, which is translated into *"Good", "Pass", or "Fail"*, based on the mean, and standard deviation of the students that participated in the same test moment as a relative standard. The scores of the four PTs in every academic year are combined, and translated into a summative decision, followed by the awarding of two study credits (of the in total 60).

After each PT, students can check their answers with an online answer key. For each answer a source is provided for further information, and for some answers a short explanation is given. Additionally, students receive their score and feedback via e-mail (*Appendix 1 – Supplemental Table 1*), and they can access feedback in ProF in the form of a table displaying their individual score, stratified by category and discipline, compared to the overall score of their peers. In ProF, their individual longitudinal test results are visualized in graphs as well [25]. There is no option to download the feedback displayed in ProF. Students receive information about the PT, and the use of ProF through a lecture in each of the bachelor years. Reflection on the feedback with their supervisor is optional.

Due to the COVID-19 restrictions, some of the PTs in the LUMC during the academic years 2020–2021 and 2021–2022 were taken from home by students, via a digital assessment platform. As the COVID-19 restrictions intensified during the pandemic (e.g. total lockdown), exam conditions varied as well. Some of the online PTs used online proctoring software, and were summative (e.g., PT1 and PT2 in 2021–2022). However, in February 2021 (PT3), we could not access the online proctoring system due to logistic reasons. Part of the students could take the PT in the exam hall, but its capacity was largely reduced due to COVID-19 regulations. As a result, the exam hall could only harbour one cohort, i.e. the third-year students. Second-year students took the PT from home, online and non-proctored. As a result, the PT was *summative* for third-year students, and *formative* for second-year students. *Figure 1* shows which PTs were formative, and summative for these two cohorts. We show the situation for these two cohorts only, because these cohorts are our main focus. Students were instructed to take the *formative* PT as a usual (proctored) PT, without using study materials, but without proctoring, we could not verify if students followed these instructions. Participation in these PTs was mandatory, but the test results were not taken into account for the rewarding of study credits. Therefore, these non-proctored PTs turned into purely formative assessments. Hereafter we will call this PT the *formative* PT, whereas the proctored PT that counts towards study credits will be called the *summative* PT.

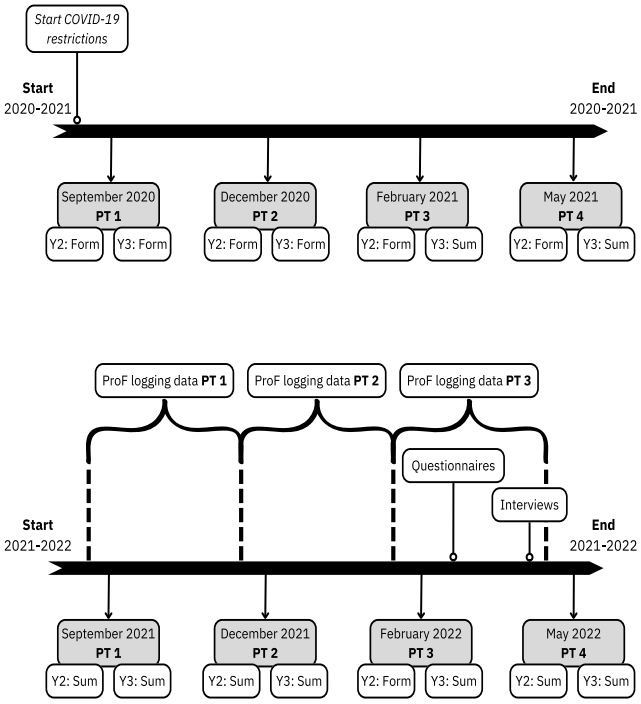


Figure 1. Timeline of progress tests and associated data collection during the academic year 2020-2021 (top) and 2021-2022. PT = progress tests; Y2 = second-year students; Y3 = third-year students; Form = formative; Sum = summative. In 2021-2022, PT3 was purely formative for Y2 students because there was no access to online proctoring, and summative for Y3 students.

Participants

All second-, and third-year bachelor medical students at the LUMC who participated in the PT session on February 2, 2022 (PT 3 of 2021–2022), were eligible for participation in the questionnaire part of the study, and all second- through six-year medical students at the LUMC were eligible for the interviews. The PT session on the February, 2, 2022 was purely formative (*formative* PT) for second-year students, while the result of the PT was taken into account for study credits in third-year students (*summative* PT). Inclusion criteria for semi-structured interviews were (1) participation in at least four of the six PTs between September 2020, and December 2021, and (2) participation in both a *formative* and *summative* PT. In total, 1286 students met our inclusion criteria. Students were sampled using maximum variation sampling based on ProF logging sessions, study-year, and PT results to ensure the representation of multiple perspectives [40]. Sampling of these students was informed by quantitative data, such as ProF logging sessions and study-year. The groups for the number of ProF logging sessions were based on the distribution among all students who participated in the PTs. The PT results were divided in two groups: *"fail"* or *"pass/good"*. If a student had failed on at least one PT, the student was assigned to the *"fail"* group (*n*=410). The other students were assigned to the *"pass/good"* group (*n*=876). We initially approached 140 students that met our sampling strategy, of whom 18 were interested. The distribution

of these students was a good representation of our sampling groups, so we invited all 18 students for an interview. After this initial sampling, second-year students were still underrepresented compared to third-year students, so we decided to sample and approach additional second-year students. Three students replied, who were all included in our study. In total, this resulted in 21 interviews, and a more equal distribution among second- (*n*=6) and third-year students (*n*=8).

Data collection

Questionnaire and ProF logging data

A questionnaire was completed either digitally or on paper (*Appendix 2 – Questionnaire*). It measured perceived assessment condition, test preparation, feedback consultation, and active use of feedback. *Perceived assessment condition* was measured with two MCQs ('formative, summative or don't know', and 'high, intermediate or low stakes'). These items were added to compare true assessment conditions with perceived assessment conditions. *Preparation* and *feedback consultation* after the PT were measured with two yes/no questions. Students were also asked to select explanatory reasons for their answers. *Active feedback use* was measured with the Active Use of Feedback (AUF) scale. This scale consists of seven 6-point, positively packed Likert-items, and is part of the validated, revised version of the Students Conceptions of Feedback (SCoF) Questionnaire [6]. Six of the seven original AUF scale items, and one item of the *'Enjoyment'* (ENJ) subscale were used. The items were adapted to the context of the PT (e.g., 'tutor' was replaced by 'progress test'). Two items were excluded because they did not apply to the specific context or were very similar to another item. The items were translated to Dutch using a forward-backward translation method. The content, and structure of the questionnaire were assessed by three master students using a thinking aloud method. Two weeks after the PT scores and feedback were made available, students received the digital questionnaire by e-mail. We also visited lectures, and working groups to hand out paper questionnaires. The students received up to two digital reminders. Age, PT grades, and ProF logging data of all students (both responders and non-responders of the questionnaire) were derived from the university's student administration system.

Interviews

We developed an interview guide to explore which factors affect feedback use in progress testing (*Appendix 3 – Interview guide*). The interview data were part of a more comprehensive study on factors influencing feedback use in progress testing [41]. In this study, we only selected interview data about students' perceptions of feedback use in the context of a formative and summative PT. Besides their own perceptions, we asked students to reflect on the ProF logging data from all bachelor, and master students in relation to formative and summative PTs during the COVID-19 pandemic (*Appendix 4 – Supplemental Figure 1*).

The principal investigator (EvW) conducted two pilot interviews with fourth-year medical students, which resulted in minor revisions in the interview guide to improve clarity and structure. The pilot interviews were not included in the study. EvW conducted the interviews with 21 students (*Appendix 5 – Supplemental Table 2*) via online meetings in Microsoft Teams in April and May 2022. Participants were invited by e-mail, and received an electronic gift card in return for participation. The interviews took 30–60 min, and were audiotaped. The audiotapes were transcribed verbatim, and anonymized before analysis. The timeline of the data collection from the different sources are depicted in *Figure 1*.

Data analysis

Questionnaire

Descriptive statistics were calculated for the demographics, and perceived assessment con-ditions. Standardized mean differences (SMD) were calculated to quantify baseline group differences between the *formative*, and *summative* PT-groups, and to explore potential response bias (non-responders versus responders) [42]. Logistic regression analyses were used to study the effect of assessment condition on test preparation, and feedback consultation. Cronbach's α was calculated to assess internal consistency of the AUF scale items. Differences between the formative, and summative PT-group were assessed by an unpaired t-test (total mean score on the AUF scale items), and chi-squared tests (multiple-choice questions on preparation, and feedback consultation). Subgroup analyses were performed on students from whom the perceived assessment condition (formative or sum-mative) matched the actual assessment condition. We used the actual assessment conditions for our main analyses, because students were well aware of the physical difference in tests condition (i.e., from home without any webcam observation versus in an exam hall with continuous supervision), and therefore we assumed that this would be a more important discriminative factor than the perceived formative or summative test condition.

ProF logging data

All ProF sessions were included for analysis, independent of the number of pageviews or duration of their session. The average number of ProF sessions per student was calculated for the PTs in September 2021 (PT 1), December 2021 (PT 2), and February 2022 (PT 3). We chose a time range of one week before the PT until one week before the subsequent PT to assess both feedback consultation before (for preparation) and after the PT. Linear regres-sion was used to estimate the effect of assessment condition on average number of ProF ses-sions for the PT in February 2022, adjusted for ProF-sessions on previous PTs (December 2022, and September 2021). Adjustment for ProF-sessions in December 2022, and September 2021 was done by adding the number of ProF-sessions around these PTs as two separate covariates in our linear regression formula. To cross-check the self-reported ProF consultation after the PT on the questionnaire, we analysed the ProF logging data of the responders in the week of the PT in February 2022 until the end of the questionnaire administration (6 weeks later). For both the questionnaire, and ProF logging data analysis, statistical significance was determined by a 95% confidence interval (CI) and $p<0.05$. Data were analysed using R version 4.1.0 (R Foundation for Statistical Computing, Vienna, Austria).

Interviews

Data analysis started after four interviews, which led to small adjustments in the interview guide to specify the questions more. The remainder of the data analysis took place after all interviews were completed. Because the extensive literature on feedback use can be integrated by *a priori* themes that guide the deductive analysis, we used template analysis in which hierarchical coding and development of successive coding templates is used [43]. Our *a priori* themes were based on EVT in the context of TTM [18, 19, 22]. Two independent coders (EvW and FvB) coded interviews 1–6 in Atlas.ti. This was discussed afterwards together and with a third researcher (AL) to reach consensus on the initial template, which was then used to guide the coding of the next interviews. Analysis of interviews 7–14 was used to further revise the initial template (EvW and FvB) which in turn was used to code interviews 15–21, and develop the final template. Only minor revisions were made to the revised initial template, and no new themes related to

the research question raised in the development of the final template, indicating theoretical sufficiency after interview 14 [44, 45]. The final template was discussed with the research team (EvW, FvB, AL, JB). During the iterative process, elements of the EVT and goal-orientation theory were incorporated in the template. Eventually, EvW reread, and recoded all interviews with the final template to ensure all relevant information to answer the research question was included in the template. With this final template, a thematic-map was constructed to identify connections between the themes and codes. Member checking was done using the Synthesized Member Checking (SMC) method [46], and yielded no adjustments.

Reflexivity

We considered and discussed (inter)personal reflexivity throughout our data collection, and analysis process using a reflective diary and critical dialogues regarding our interpretations of the data [47]. The reflective diaries created awareness of personal expectations, assumptions, and reactions to the participants and data, and were used to guide the dialogues between the investigators. In interviewing the students, EvW experienced that she could easily relate to the participants, because of her own medical background and experience with the PT. This created an open atmosphere, in which the students felt comfortable to talk openly about their experiences and perceptions. Influenced by her scientific background in (bio)medicine, EvW attempted to attain as much objectivity and produce rigorous qualitative research by using maximum variation sampling, member checking, and reflexivity throughout the data collection and analysis. The other researchers were an educational consultant and researcher in medical education (FvB) and a medical doctor with experience in clinical teaching and educational research (AL). FvB has been trained to conduct research in an empirical way during his studies in cognitive psychology. As such, he supported using theoretical concepts from feedback literature to formulate *a priori* themes. This theory-driven approach may have influenced the results. AL is a member of the national PT working group and a PT examiner, which might have led to assumptions on study behaviour based on her experience with the PT and conversations with students in the past. Her involvement with the PT was very valuable in reflecting on the interview data, and placing it in the right context.

Results

Demographics and perceived assessment condition

Of 316 students who took the purely formative PT (*formative* PT-group), 113 students participated in the questionnaire (response rate: 35.8%). In the *summative* PT-group, 154 students participated in the questionnaire (response rate: 50.0%) from which 3 students were excluded due to incomplete reply to the questionnaire (*Appendix 6 – Supplemental Figure 2*). Responders (*n*=264) and non-responders (*n*=354) differed in fail/pass/good grade and average ProF logging sessions (mean (SD); 1.29 (1.60) versus 0.71 (1.87), for responders versus non-responders) (*Appendix 7 – Supplemental Table 3*). In both the formative and summative PT-group, 70% of the responders were female, and the distribution of the grades was similar among the groups (*Table 1*). Regarding the perceived stakes of the PT in February 2022, 50% versus 13% of students perceived the PT as low stakes, 42% versus 62% as intermediate stakes, and 7% versus 25% as high stakes for *formative* and *summative* PT-group respectively (*Appendix 8 – Supplemental table 4*). The perceived assessment conditions formative and summative can also be found in *Appendix 8 – Supplemental Table 4*.

Table 1. Baseline characteristics of the responders of the questionnaire in the *formative* and *summative* progress test group.

	Overall (*n*=264)	Formative Test (*n*=113)	Summative Test (*n*=151)	SMD[a]
Age, median (IQR)	21 (20, 21)	20 (20, 21)	21 (21, 22)	0.841
Female, n (%)	185 (70)	80 (71)	105 (70)	0.022
Grade, n (%)				
Fail	24 (9)	11 (10)	13 (9)	0.034
Pass	106 (40)	46 (41)	610 (40)	0.020
Good	134 (51)	56 (50)	78 (51)	0.040
Proportion passed earlier PTs, %				
Sep '21[b]	223 (87)	99 (88)	124 (87)	0.030
Dec '21	226 (87)	97 (86)	129 (88)	0.059

SMD = standardized mean difference; IQR = interquartile range; PT = progress test.

[a]A standardized mean difference >0.1 may point towards meaningful imbalance between groups.

[b]PTs of Sep '21 and Dec '21 were summative tests

In the following paragraphs we present the results for each research question: the effect of a *summative* PT and a *formative* PT on medical students' (1) test preparation (questionnaires, and interviews), (2) factors that influence test-taking motivation, and the use of feedback (interviews), and (3) self-reported and actual feedback use after the test (questionnaires and ProF logging data, and interviews).

Test preparation

Logistic regression showed no significant association between assessment condition and preparation for the PT (adjusted OR [aOR] 1.26, 95% CI 0.57–2.76) (*Table 2*). A similar result was found in the subgroup analysis (aOR 1.83, 95% CI 0.72–4.64).

Table 2. Test preparation, feedback consultation and active use of feedback of students in the *formative* and *summative* progress test-group.

	Formative Test	Summative Test	Crude OR (95% CI)	Adjusted OR (95% CI)[a]	p-value
True formative and summative					
Number of individuals	113	151			
Preparation, n (%)	14 (12)	28 (19)	1.61 (0.80-3.22)	1.26 (0.57-2.76)	0.568
Feedback consultation, n (%)					
Answer key	22 (19)	56 (37)	2.44 (1.38-4.32)	1.92 (1.04-3.55)	0.038
Feedback e-mail	89 (79)	126 (83)	1.36 (0.73-2.53)	1.00 (0.49-2.05)	0.996
Feedback ProF	41 (36)	86 (57)	2.32 (1.41-3.83)	1.92 (1.10-3.34)	0.021
None	20 (18)	13 (9)	2.28 (1.08-4.81)	1.86 (0.80-4.32)	0.149
ProF logging data[c], n (%)	26 (23)	58 (38)	2.09 (1.21-3.61)	1.89 (1.03-3.44)	0.039
Number of individuals[d]	90	135	**t-value**	**95% CI**	
Active use of feedback, mean (SD)	3.2 (0.9)	3.1 (0.9)	1.09	-0.10-0.36	0.275[e]
Perceived formative and summative[b]					
Number of individuals	79	128			
Preparation, n (%)	8 (10)	25 (20)	2.15 (0.92-5.05)	1.83 (0.72-4.64)	0.205
Feedback consultation, n (%)					
Answer key	18 (23)	48 (37)	2.03 (1.08-3.84)	1.47 (0.73-2.94)	0.280
Feedback e-mail	60 (76)	106 (82)	1.53 (0.76-3.04)	1.07 (0.48-2.39)	0.876
Feedback ProF	26 (33)	74 (57)	2.79 (1.55-5.02)	2.25 (1.18-4.31)	0.014
None	15 (19)	11 (9)	2.49 (1.08-5.75)	1.94 (0.74-5.05)	0.175
ProF logging data[c], n (%)	17 (22)	48 (38)	2.19 (1.15-4.17)	1.80 (0.88-3.66)	0.106
Number of individuals[d]	62	114	**t-value**	**95% CI**	
Active use of feedback, mean (SD)	3.2 (0.8)	3.1 (0.8)	0.94	-0.14-0.38	0.351[e]

[a]Adjusted for age and result progress test December 2021 (fail, pass, good).

[b]Subgroup analysis; Perceived formative/summative = students in the purely formative/summative test group who knew it was formative/summative.

[c]Real-time ProF logging data in week 05 (PT administration) until week 11 (end of questionnaire administration).

[d]Students who consulted feedback in e-mail or progress test feedback system.

[e]Unpaired t-test.

Regarding the reasons why students did not prepare for the PT, 27% of the students in the *formative* PT-group stated on the questionnaire that the PT was not important compared to 1% of the students in the *summative* PT-group ($p<0.001$, *Appendix 9 – Supplemental Table 5*). In the subgroup analysis this difference became more prominent (26 (37%) versus 0 (0%), $p<0.001$ for *perceived formative* versus *perceived summative*). Other reasons for not preparing were a lack of consequences and not knowing how to prepare.

In the interviews, many students mentioned that the lack of consequences and the possibility to look up answers in the formative PT affected their test preparation:

"My preparation for a formative PT is worse. I still look up some things in advance which I just want to know, but there is less pressure, so if it does not work out or if I don't really feel like doing it, then I think, well, if a question comes up I don't know, I can just look it up." (Interview #3)

Factors that influence test-taking motivation and feedback use

The value given to the *formative* and *summative* PT influenced students test-taking motivation, and determined how students behaved during the formative PT (i.e., test-taking behaviour). The majority of students valued the *summative* PT as more important compared to the *formative* PT, because of its' consequences for study progress, and the more formal test-setting compared to the *formative* PT (on location vs. at home). We call these students 'performance-oriented' (*Figure 2*, upper path):

"Ultimately, you take each test for the study credits. You follow the lessons to learn something, but I do not make a test to learn from it. I make a test to see if my learning was successful. And whether or not I can receive the credits so I can continue." (Interview #4)

On the other hand, 'learning-oriented' students valued the test and its' feedback as a moment of self-assessment and reflection, regardless of the assessment condition. Their main focus in both the *formative* and *summative* PT was to assess their current knowledge level, gain insights in their own strengths and weaknesses and learn from what they did wrong (*Figure 2*, lower path).

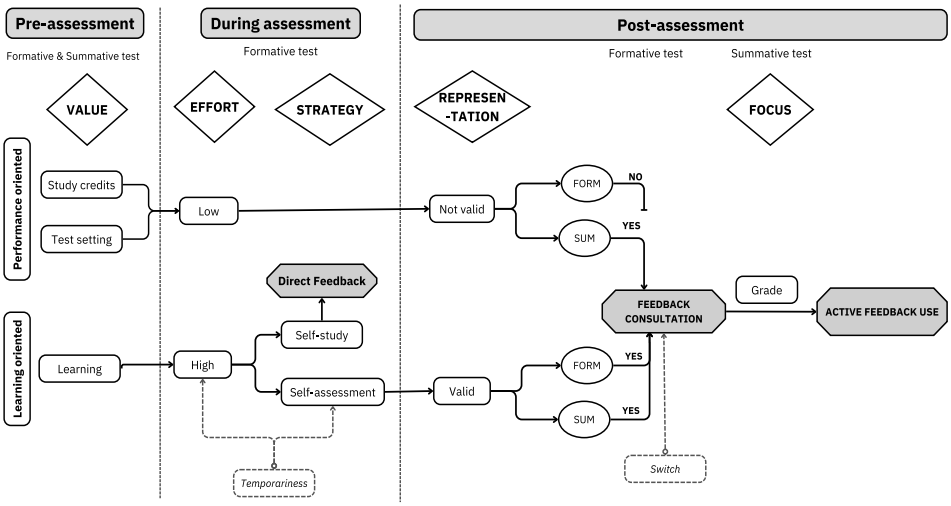


Figure 2. Thematic map showing the connections between the themes (triangular shapes on the top) and codes (boxes below the themes) pre-, during, and post-assessment for performance-oriented (upper path) and learning-oriented students (lower path). Form = formative progress test; Sum = summative progress test.

Test-taking behaviour: effort and strategy

We distinguished two subthemes within test-taking behaviour: effort and strategy. These themes only relate to the *formative* PT, because the low-stakes and lack of supervision were perceived as an opportunity to adapt their test-taking behaviour according to their values and goals in relation to the PT.

Learning-oriented students tended to put significant effort in the *formative* PT, as they wanted to be able to reflect effectively on their performance. In contrast, performance-oriented students put less effort in taking the *formative* PT, reflected by a higher proportion of guessing, looking up answers on the internet or being less focused during the test:

"I think that I guessed more of the answers in the online (formative) test when I recognized an answer vaguely from a previous course. I did not know the answer completely for sure, but I was doubting between three options and then I just guessed because it did not matter so much." (Interview #7)

Students employed different test-taking strategies in the *formative* PT, which could be divided in self-study and self-assessment. The self-study strategy was characterized by using study materials to look up answers during the test, mainly with the idea to learn directly from it. By looking up answers for questions, they generated instant feedback for themselves and hence used the *formative* test as a guide for self-study:

"Well, I thought if I look it up right away I will learn something from it, because then I know the answer. And if I will not look at it anymore afterwards, then I actually do not learn so much either, because I don't know if my answers were correct or incorrect." (Interview #9)

In the self-assessment strategy, students approached the *formative* PT as if it were a *summative* PT and refrained from looking up answers. They used the test as a realistic self-assessment of their current knowledge:

"When you get the result, that you have some sort of measurement of how good you actually are at it. Because otherwise (when using study material) I have the idea that it does not make sense at all to take that test." (Interview #10)

Contextual factors: Temporariness

Many students took into account that the *formative* PTs were only temporary and that in the near future, they would become summative again. This temporariness encouraged them to make the *formative* test just as seriously as the *summative* test, with an indirect focus on study credits relating to the performance-oriented mindset:

"Of course I could have looked it all up, but then I think you will fall at a certain moment. I think you cannot sustain that when the test is proctored again. And then it's only annoying that you're going to drop in your score again." (#Interview 14)

Self-reported and actual feedback use after the test

Feedback consultation

Students who took the *summative* PT reported consulting ProF (aOR 1.92, 95% CI 1.10–3.34) and the answer key (aOR 1.92, 95% CI 1.04–3.55) more often than students who made the *formative* PT. In *perceived formative* versus *summative*, the effect on feedback consultation in ProF became more evident (aOR 2.25, 95% CI 1.18–4.31) and the adjusted effect on the answer key consultation was not observed

(aOR 1.47, 95% CI 0.73–2.94) (*Table 2*). The sensitivity analysis with real-time ProF logging data of the responders showed the same trend as the self-reported data, but yielded lower overall numbers (26 (23%) versus 58 (38%), aOR 1.89, 95% CI 1.03–3.44; for *formative* versus *summative*; 17 (22%) versus 48 (38%), aOR 1.80 (0.88–3.66); for *perceived formative* versus *perceived summative*) (*Table 2*). Besides the sensitivity analysis using only ProF logging data of the responders, we also analysed the ProF logging data of all participating students including the non-responders. This analysis showed that there were more ProF logging sessions around the *summative* PT in February 2022 than around the *formative* PT (β:0.444, *p*<0.001). After adjustment for logging behaviour in earlier *summative* PTs in September 2021 and December 2022, this effect remained significant (β:0.251, *p*:0.003) (*Figure 3, Appendix 10 – Supplemental Table 6*).

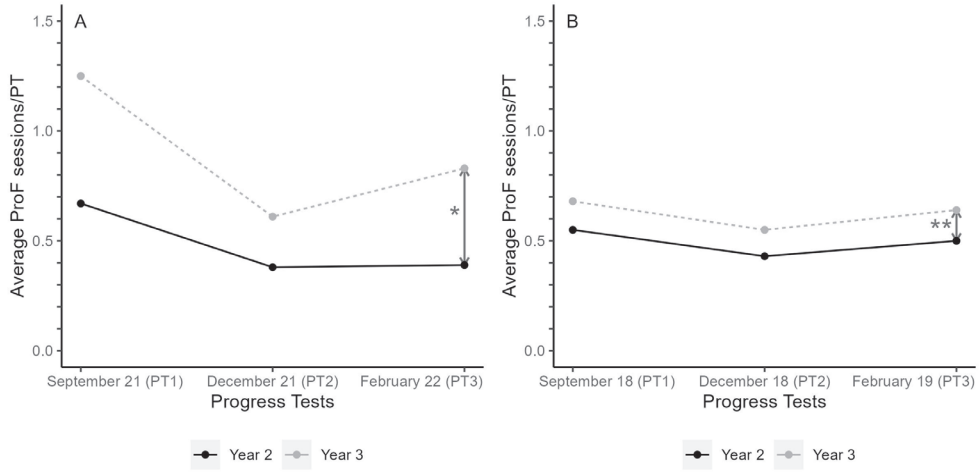


Figure 3. Average ProF sessions in year 2 (black line) and year 3 (dotted grey line) for the progress tests in September 2021, December 2021, and February 2022. Each point on the curve represents the average number of ProF sessions per student; *: crude beta: 0.444, p<0.001; adjusted beta: 0.251, p: 0.003.

Students who reported not to consult the feedback after the *formative* PT on the questionnaire more often considered the test as not important (15 (22%) versus 1 (2%), *p*<0.001 for *formative* versus *summative*) (*Table 3*). This resonates with the perceptions of the performance-oriented students:

"I think that you are more motivated when the test counts for study credits, so then afterwards you will be more interested in how you performed because it counts." (Interview #10)

However, qualitative data also revealed the learning-oriented students who valued the feedback of both assessment conditions for their learning (*Figure 2,* lower path):

"I look at the test result to know what questions I did wrong and to learn from it. And it doesn't matter to me whether it is formative or summative, because that remains the same. I still want to know which questions I got right and wrong. And I still want to know, I still want to learn from the things I did wrong. So, then it doesn't matter if the test was formative or summative." (Interview #3)

In the questionnaires, the *summative* PT-group more often found the feedback not useful (1 (1%) versus 8 (13%), *p*:0.015 for *formative* versus *summative*). Similar results were found for perceived assessment conditions (*Table 3*). Other reasons for not consulting feedback included no awareness of or not understanding ProF, not knowing how to use the feedback, and a lack of interest.

Table 3. Reasons for not using the progress test feedback system in the *formative* and *summative* progress test group.

	Formative Test	Summative Test	p-value[a]
True formative and summative			
Number of individuals	67	63	
No ProF use, n (%)			
Findability[c]	25 (37)	14 (22)	0.061
Time	24 (36)	16 (25)	0.178
Effort	7 (10)	1 (2)	0.062[d]
Importance	15 (22)	1 (2)	0.000
Grade	33 (49)	22 (35)	0.083
Utility	1 (1)	8 (13)	0.015[d]
Answer key	3 (4)	7 (11)	0.200[d]
Other	4 (6)	12 (19)	0.025
Perceived formative and summative[b]			
Number of individuals	48	52	
No ProF use, n (%)			
Findability	20 (42)	12 (23)	0.055
Time	16 (33)	15 (29)	0.628
Effort	5 (10)	1 (2)	0.102
Importance	13 (27)	1 (2)	0.000
Grade	21 (44)	20 (38)	0.591
Utility	0 (0)	6 (12)	0.027[d]
Answer key	1 (2)	5 (10)	0.207[d]
Other	3 (6)	10 (19)	0.054

ProF = progress test feedback system.
[a]Chi-squared test.
[b]Subgroup analysis; Perceived formative/summative = students in the formative/summative test group who knew it was formative/summative.
[c]Findability: ***"I do not know where I can find the feedback"***; Time: ***"I did not have time to look at the feedback"***; Effort: ***"I did not put effort in this progress test"***; Importance: ***"I thought this progress test was not important"***; Grade: ***"I got a pass/good for this progress test"***; Utility: ***"I find the feedback not useful"***; Answer key: ***"I already checked my answers with the answer key"***.
[d]Fisher's exact test

Qualitative data revealed that representation of the *formative* test also played a role in feedback consultation (*Figure 2*). Performance-oriented students indicated less interest in the feedback of the *formative* PT, because their low test-taking effort in taking the formative PT did not provide a valid representation of their own knowledge level. Therefore, the feedback was less meaningful to them:

"I think I took a quick look at ProF. That I just looked at that line, but that I thought yes, it is probably now higher than it should be. So I did not attach much value to it." (Interview #4)

This was also the case for students who used study material during the *formative* PT. Besides, these students found it more useful to receive direct feedback during the PT. In contrast, for learning-oriented students who used the *formative* PT as self-assessment the test result was a valid representation,

and they were interested to consult the feedback to assess their strengths and weaknesses.

ProF consultation was relatively high after the first *formative* PT (September 2020, *Appendix 4 – Supplemental Figure 1*). The interviewed students mentioned that this could be explained by curiosity right after switching to formative testing (*Figure 2, 'Switch'*):

"The first time it is always exciting, oh new and what would be my result now that it's online for the first time. And is there a difference maybe with the paper version that I always had. So then it is a bit more interesting and if you've done a few then you just think oh it's going fine, whatever." (Interview #18)

Active feedback use

The internal consistency (Cronbach's α) of the 6-point Likert scale items was 0.85 (>0.80: acceptable) [48, 49]. After deletion of item 7 of the subfactor ENJ the Cronbach's α remained 0.85. We found no difference in the mean total score on the items of AUF and ENJ (3.2 (0.9) versus 3.1 (0.9), t(223): 1.09, 95% CI -0.10-0.36). Comparing perceived assessment conditions yielded the same result (*Table 2*). On item level, item 7 (enjoy) had the highest score (5 (4–6), whereas item 3 (setting goals) and 6 (changing learning) had the lowest score (2 (1–3) in both groups (*Appendix 11 – Supplemental Table 7*).

Students who were interested in the test result and feedback often only consulted the feedback without actively using it. They seemed to use the feedback as a 'thermometer' to assess if they were still at the right 'temperature'. If they were still on the right track, they did not feel the urgency to change anything and engage with the feedback: *"If it ain't broke, don't fix it"(# Interview 4).* An insufficient grade on the other hand was or will be an incentive to act on the feedback and use it to prepare for the next PT:

"Suppose, if I had failed I would think oh dear, then I will really look at what I did wrong, which subject and really do that because you still want to get those study credits." (Interview #18)

Although the formative PT was also graded, grade focus only occurred in the *summative* PT, mainly because an insufficient grade on the *formative* PT had no consequences. Thus, no need was felt to act on the feedback (*Figure 2*).

Discussion

In this mixed-methods study, we compared the effect of a purely formative PT (*formative* PT) with a PT with a summative component (*summative* PT) on medical students' feedback use and test-taking motivation. We triangulated quantitative and qualitative interview data to explain these in the context of a *formative* versus a *summative* PT. Our thematic map (*Figure 2*), based on our qualitative data, in which EVT and goal-orientation frameworks were integrated helped explain our quantitative results, and provided a nuanced picture of the different ways students approached the feedback in the *formative* and *summative* PT. Test preparation was relatively low for both PT assessment conditions and did not differ between groups. Qualitative data showed that test-taking motivation and feedback use relate to how students value the assessment. Performance-oriented students valued the *summative* PT as more important because of its' consequences for study progress, and learning-oriented students valued the PT feedback for their own learning, regardless of the assessment condition.

These orientations influenced their test-taking behaviour (effort and strategy), and feedback consultation (representation of *formative* PT results). Self-reported questionnaire data showed more ProF consultation and use of the answer key after the summative PT compared to the *formative* PT. Actual feedback use, measured by ProF logging data, showed the same results. Students in the *formative* PT-group who did not consult PT feedback more often reported the *formative* PT as unimportant, reflecting the perceptions of performance-oriented students. However, self-reported active feedback use after the PT was relatively low in general and did not differ between groups, which was mainly determined by grade focus.

Test preparation, feedback consultation and test-taking motivation

We measured test preparation and feedback consultation with different data sources. The ProF logging data demonstrate that, in general, students made limited use of ProF to consult feedback, which is important to take into account with the interpretation of our data. Despite low use of ProF, our data can contribute to a better understanding of feedback behaviour. Both questionnaire and interview results suggest that the motivation to prepare and consult feedback relates to how students value the assessment. The interview results revealed that experienced utility value (i.e. usefulness) and attainment value (i.e. importance) of the PT affected test-taking effort, the important component of test-taking motivation, which influenced feedback consultation [22]. This positive relation between value and effort has also been found in test-taking motivation with test performance as outcome [19, 23, 50].

Moreover, our interview data showed that students valued the different PT assessment conditions based on whether they were orientated towards performance or learning. This aligns with the goal-orientation theory, which states that performance-orientated students focus on achievement based on normative standards (i.e. study credits), whereas learning-orientated students focus on achievement based on learning [24]. It seems that students' goal orientation guided their test effort and engagement with the feedback. Although students did not explicitly state goals for the PT in our study, they did show a more general focus on either learning or performance. Below, we elaborate on students' performance and learning orientation in this study.

Performance-oriented students

Performance-orientated feedback behaviour was revealed by our qualitative interview data, and confirmed by our quantitative results. The interviews showed that students found the summative PT more important and valuable, which led to less test-taking effort and feedback consultation after the *formative* PT. Quantitative data confirmed this performance orientation as self-reported, and actual feedback consultation was higher after the *summative* PT. Also, the perception that the PT was not important, and thus ProF consultation or preparation was not needed, was more profound in students who participated in the *formative* PT. These results are visualized in the upper path of our thematic map (*Figure 2*). The performance-oriented students mainly focus on the direct personal consequences (i.e., study-credits) of the test in the pre-assessment phase, see no value in investing effort in the *formative* tests (i.e., low effort during assessment), leading to an invalid representation of their test result, and a decreased motivation to consult the feedback of the *formative* test post-assessment. The study credits in the summative test, on the other hand, motivated these students to consult the feedback.

Learning-oriented students

Our qualitative analysis suggested that students were not also focused on learning. Thus, the interview data further deepened and nuanced our understanding of students' feedback use. Learning-orientated students valued the PT and its feedback as part of their learning process, regardless of the assessment condition (*Figure 2,* lower path). These students took the *formative* test seriously, invested high effort, used it for self-assessment or self-study, and were motivated to consult the feedback of the *formative* test. This aligns with previous research in surgical residents showing that formative assessments promoted a learning-oriented motivation [51]. The strategy of self-study was interesting in that the test itself was used as tool to pay attention to knowledge gaps and generate direct feedback. It is more likely that these students benefit from formative assessments and engage in more self-regulated learning compared to students adopting the performance orientation [52-54].

Some learning-oriented students indirectly also focused on the study credits considering that the *formative* PTs would switch back to *summative*. *Figure 2* shows that this contextual factor (*'Temporariness'*) influenced test-taking effort, and strategy in the *formative* test of these students (with the dotted arrows). They decided to put high effort in the *formative* test, and use it as self-assessment, to make sure they were at the right level to pass the upcoming summative test. Although these students predominantly focus on learning in the *formative* PTs, they do not completely let go their performance-orientation for the study credits of the future *summative* PTs.

Active use of feedback

Besides students' feedback consultation in the e-mail, ProF or by using the answer key, which can be considered a more passive use of feedback, we also measured active use of feedback after the PT by the AUF scale items in our questionnaire [6]. Although students enjoyed receiving feedback, active use of feedback after the PT was relatively low and no difference was found between the groups. The interview data also showed that most students did not actively use the feedback, and that they tended to act only on the feedback when they failed on the *summative* PT. This is illustrated in *Figure 2* in the post-assessment phase, where all students, regardless of their orientation, were driven by the grade in their decision to act on the feedback after consultation. This suggests that failure drove using feedback, regardless of students' learning orientation. However, we could only find qualitative evidence for this, as too few students failed the PT to provide quantitative evidence. As described in previous literature, grade focus strongly limits the likelihood to engage with feedback after a sufficient summative grade [55, 56]. However, students stated even less engagement with the feedback after the *formative* PT, because they lacked a feeling of urgency to change something as this test had no direct consequences for their study progress. These findings are in line with earlier studies on progress testing, where the effect of the feedback on learning was questionable [26, 32, 33, 35, 57]. Although students used the feedback to monitor their progress, and identify strengths and weaknesses in these studies, there was no direct influence on future learning [32, 33].

Implications for practice

Our results suggest that the desired positive effect of formative testing on the learning process is limited in progress testing, with students mainly focusing on performance. Introducing more formative assessments in medical education requires a change in shift in focus towards the learning process (learning-oriented)

rather than the outcome (performance-oriented), and enhancing students' feedback literacy: their ability to effectively engage with and utilize feedback [58]. This involves creating a supportive environment in which students are encouraged to develop feedback literate skills [59]. An example of such a system emphasizing the value of assessment for learning is the programmatic assessment approach. In this approach assessments are no longer divided in formative and summative, but rather represent a continuum of stakes (from low to high). Heeneman *et al.* demonstrated positive results on feedback use of embedding a formative PT in a programmatic assessment system, in which the reflection on the PT, and guidance in the feedback process by mentors in the curriculum is embedded [28]. A supportive assessment environment that emphasizes the understanding of the concept, and purpose of formative testing is key in motivating students and support learning [60, 61].

Strengths and limitations

In the present study we had the unique opportunity to make a direct comparison between two conditions of the same test in one medical curriculum. Except for the assessment conditions, the educational setting was exactly the same for all students and feedback was provided to all students, which facilitated the assessment of the (additional) effect of the summative component over the formative component of assessment on feedback use. Moreover, we analysed both assigned and perceived test conditions, which showed the same trend. Additionally, triangulation of quantitative, and qualitative data was used to increase validity and create a more in-depth understanding of student's values. The triangulation of three data sources also adds to the credibility of our conclusion that the formative PT was associated with less feedback use than the summative PT.

This study also has some limitations. Firstly, this study was conducted at only one medical school, which could limit the transferability to other settings. Secondly, we cannot completely rule out that the difference in study-years between the groups affected feedback behaviour. As third-year students are more experienced with the PT, possibly having a more serious attitude towards their study, this might have resulted in a higher baseline level of feedback use. Nevertheless, test preparation was similar between groups, and the effect found in the ProF logging data remained significant after adjusting for previous ProF use in both years. Moreover, the interview data clarified that the formative and summative component of the PT played a significant role in their feedback behaviour, regardless of their study progress. Thirdly, the responders to our questionnaires were overall students with more ProF logging sessions, and the response rate of the students in the formative PT-group was relatively low. However, the ProF logging data of all students, both responders and non-responders, point towards the same conclusion that feedback consultation was higher after the summative PT. Fourthly, the assessment of a more longitudinal pattern of ProF logging behaviour under summative conditions was hindered by changes in PT conditions before September 2021 (COVID-19) and after February 2022 (new adaptive format). Finally, it must be noted that this study focused on the PT, which is a longitudinal, repetitive and comprehensive assessment. We cannot be sure to what extent these results can be adapted to other contexts, such as a context with a different assessment structure or to end-of-course examinations. The perception of feedback and the feedback behaviour in these other contexts is an interesting question for future research. Moreover, additional research is needed to understand the interaction between the different goal orientations and feedback use.

Conclusion

In conclusion, this study found that students make little use of PT feedback. When they do use PT feedback, a *summative* PT is associated with more feedback consultation compared to a *formative* PT, which can be explained by lower overall test-taking motivation in the *formative* PT and a performance-orientation. Nonetheless, qualitative data also showed learning-oriented students who found the *formative* PT useful and important for their learning, emphasizing that the perceived value of assessment is key to the learning effect of formative testing. Active use of feedback after the PT was low in both assessment conditions and seemed to be affected mostly by high-stakes consequences (i.e., not obtaining enough study credits due to failing the *summative* PT). This might be partly because reflection, and guidance in the feedback process were not embedded in the curriculum. Therefore, it is important to consider the introduction of formative assessments in the medical curriculum very carefully, and make sure students understand its value and are supported in the feedback process.

References

1. 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.
2. Newble DI, Entwistle NJ. Learning styles and approaches: implications for medical education. Medical Education. 1986;20(3):162-75.
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. 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.
5. Black P, Wiliam D. Assessment and Classroom Learning. Assessment in Education: Principles, Policy & Practice. 1998;5(1):7-74.
6. 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.
7. 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.
8. Koh LC. Refocusing formative feedback to enhance learning in pre-registration nurse education. Nurse Education in Practice. 2008;8(4):223-30.
9. Kulasegaram K, Rangachari PK. Beyond "formative": assessments to enrich student learning. Advances in Physiology Education. 2018;42(1):5-14.
10. Schuwirth LWT, van der Vleuten CPM. The use of progress testing. Perspectives on Medical Education. 2012;1(1):24-30.
11. Scott IM. Beyond 'driving': The relationship between assessment, performance and learning. Medical Education. 2020;54(1):54-9.
12. Seligman L, Abdullahi A, Teherani A, Hauer KE. From Grading to Assessment for Learning: A Qualitative Study of Student Perceptions Surrounding Elimination of Core Clerkship Grades and Enhanced Formative Feedback. Teaching and Learning in Medicine. 2021;33(3):314-25.
13. Watling CJ, Ginsburg S. Assessment, feedback and the alchemy of learning. Medical Education. 2019;53(1):76-85.
14. Hattie J, Timperley H. The Power of Feedback. Review of Educational Research. 2007;77(1):81-112.
15. Zimmerman BJ. Investigating Self-Regulation and Motivation: Historical Background, Methodological Developments, and Future Prospects. American Educational Research Journal. 2008;45(1):166-83.
16. Nicol DJ, Macfarlane-Dick D. Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. Studies in Higher Education. 2006;31(2):199-218.
17. Wormald BW, Schoeman S, Somasunderam A, Penn M. Assessment drives learning: an unavoidable truth? Anatomical Sciences Education. 2009;2(5):199-204.
18. Baumert J, Demmrich A. Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. European Journal of Psychology of Education. 2001;16(3):441-62.
19. Cole JS, Bergin DA, Whittaker TA. Predicting student achievement for low stakes tests with effort and task value. Contemporary Educational Psychology. 2008;33(4):609-24.
20. Thelk AD, Sundre DL, Horst SJ, Finney SJ. Motivation Matters: Using the Student Opinion Scale to Make Valid Inferences About Student Performance. The Journal of General Education. 2009;58(3):129-51.
21. Wise SL, DeMars CE. Low Examinee Effort in Low-Stakes Assessment: Problems and Potential Solutions. Educational Assessment. 2005;10(1):1-17.
22. 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.
23. Zilberberg A, Finney SJ, Marsh KR, Anderson RD. The role of students' attitudes and test-taking motivation on the validity of college institutional accountability tests: A path analytic model. International Journal of Testing. 2014;14:360-84.
24. Elliot AJ, Dweck CS. Competence and Motivation: Competence as the Core of Achievement Motivation. Handbook of competence and motivation. New York, NY, US: Guilford Publications; 2005. p. 3-12.
25. 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.
26. 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.
27. 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.
28. 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.

29. 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.
30. 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):Doc41.
31. 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.
32. Aarts R, Steidel k, Manuel BAF, Driessen EW. Progress testing in resource-poor countries: A case from Mozambique. Medical Teacher. 2010;32(6):461-3.
33. 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.
34. 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.
35. 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.
36. 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.
37. Fetters MD, Curry LA, Creswell JW. Achieving Integration in Mixed Methods Designs—Principles and Practices. Health Services Research. 2013;48(6 Pt 2):2134-56.
38. Maxwell JA, Mittapalli K. Realism as a Stance for Mixed Methods Research. SAGE Handbook of Mixed Methods in Social & Behavioral Research: SAGE Publications, Inc.; 2010. p. 145-68.
39. Lord FM. Formula scoring and number-right scoring. Journal of Educational Measurement. 1975;12(1):7-11.
40. Onwuegbuzie A, Collins K. A Typology of Mixed Methods Sampling Designs in Social Science Research. The Qualitative Report. 2015.
41. van Wijk EV. Understanding students' feedback use in medical progress testing: A qualitative interview study (Manuscript submitted for publication). 2023.
42. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behavioral Research. 2011;46(3):399-424.
43. Brooks J, McCluskey S, Turley E, King N. The Utility of Template Analysis in Qualitative Psychology Research. Qualitative Research in Psychology. 2015;12(2):202-22.
44. Dey I. Grounding grounded theory: guidelines for qualitative inquiry. San Diego: Academic Press; 1999;282 p.
45. Saunders B, Sim J, Kingstone T, Baker S, Waterfield J, Bartlam B, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Quality & Quantity. 2018;52(4):1893-907.
46. Birt L, Scott S, Cavers D, Campbell C, Walter F. Member Checking: A Tool to Enhance Trustworthiness or Merely a Nod to Validation? Qualitative Health Research. 2016;26(13):1802-11.
47. Olmos-Vega FM, Stalmeijer RE, Varpio L, Kahlke R. A practical guide to reflexivity in qualitative research: AMEE Guide No. 149. Medical Teacher. 2022;0(0):1-11.
48. Lance CE, Butts MM, Michels LC. The Sources of Four Commonly Reported Cutoff Criteria: What Did They Really Say? Organizational Research Methods. 2006;9:202-20.
49. Nunnaly JC, Bernstein, I.H. Psychometric theory. 3rd ed. New York, NY, US: McGraw-Hill;1994.
50. Penk C, Schipolowski S. Is it all about value? Bringing back the expectancy component to the assessment of test-taking motivation. Learning and Individual Differences. 2015;42:27-35.
51. Lund S, D'Angelo JD, Gardner AK, Stulak J, Rivera M. General surgery resident motivation: the effect of formative compared to summative simulated skills assessments. Global Surgical Education - Journal of the Association for Surgical Education. 2022;1(1):55
52. Ames C. Classrooms: Goals, structures, and student motivation. Journal of Educational Psychology. 1992;84:261-71.
53. Pintrich PR. The Role of Goal Orientation in Self-Regulated Learning. Handbook of Self-Regulation: Elsevier; 2000. p. 451-502.
54. Schunk DH, Pintrich PR, Meece JL. Motivation in education: theory, research, and applications. 3rd ed. Upper Saddle River, N.J.: Pearson/Merrill Prentice Hall. 2008;433.
55. Harrison CJ, Könings KD, Schuwirth L, Wass V, van der Vleuten C. Barriers to the uptake and use of feedback in the context of summative assessment. Advances in Health Sciences Education. 2015;20(1):229-45.
56. 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.
57. Schüttpelz-Brauns K, Kadmon M, Kiessling C,

Karay Y, Gestmann M, Kämmer JE. Identifying low test-taking effort during low-stakes tests with the new Test-taking Effort Short Scale (TESS) – development and psychometrics. BMC Medical Education. 2018;18(1):101.

58. Molloy E, Boud D, Henderson M. Developing a learning-centred framework for feedback literacy. Assessment & Evaluation in Higher Education. 2020;45(4):527-40.
59. Carless D, Winstone N. Teacher feedback literacy and its interplay with student feedback literacy. Teaching in Higher Education. 2023;28(1):150-63.
60. Heeneman S, Oudkerk Pool A, Schuwirth LWT, van der Vleuten CPM, Driessen EW. The impact of programmatic assessment on student learning: theory versus practice. Medical Education. 2015;49(5):487-98.
61. Nouns ZM, Georg W. Progress testing in German speaking countries. Medical Teacher. 2010;32(6):467-70.

Appendix

Appendix 1 – Supplemental Table 1

Supplemental Table 1A. Feedback of the progress test with the results per category provided by e-mail.

		Individual				Test moment group (*n*=57)							
Description categories	**Number of questions**	**Correct**	**Incorrect**	**?**	**Score**	**Correct**	**Std**	**Incorrect**	**Std**	**?**	**Std**	**Score**	**Std**
01 Respiratory system	13	69	31	0	56	68	13	28	12	4	7	57	18
02 Musculoskeletal system	17	59	41++	0--	38	58	11	31	9	11	10	44	14
03 Mental Health Care	16	69	31+	0	58-	75	14	20	12	5	7	68	18
04 Reproductive system	11	45-	55++	0	27--	58	15	29	13	13	13	48	18
05 Blood, lymph, heart and circulation	24	58	25	17+	48	60	13	29	11	11	9	48	17
06 Hormones and metabolism	13	46-	46++	8	-29	57	13	31	14	12	10	46	17
07 Skin and connective tissue	12	83	17	0	78	80	10	17	10	3	6	74	13
08 Personal, social and prevention aspects	17	29--	71++	0--	4--	52	14	38	14	11	10	35	19
09 Digestive system	17	71	29	0--	61	66	12	26	11	8	7	57	15
10 Kidneys and urinary tract	16	69	25	6	59	71	13	21	11	7	8	63	16
11 Nervous system and senses	17	47--	47++	-6	28--	62	13	26	12	12	11	53	16
12 Knowledge about skills	23	48	39	13	33	49	11	40	11	11	9	32	14
Total	196	57-	38++	5-	42-	62	8	29	6	9	6	51	9

-/--/++/+ low respectively high in comparison with the total group. Results are presented in percentages. Std = standard deviation. ? = question mark option use.

Supplemental Table 1B. Feedback of the progress test with the results per discipline provided by e-mail.

		Individual				Test moment group (n = 57)							
Description disciplines	**Number of questions**	**Correct**	**Incorrect**	**?**	**Score**	**Correct**	**Std**	**Incorrect**	**Std**	**?**	**Std**	**Score**	**Std**
Anatomy	12	58	33	8	46	60	15	34	14	6	9	48	20
Biochemistry, molecular and cellular biology and genetics	18	50	44++	6--	34	46	14	31	12	24	14	34	17
Pharmacology	8	62	25	12	54	65	15	27	14	8	9	54	20
Physiology	11	73	27+	0-	62	73	17	18	12	9	12	65	21
Patho-, immuno- en microbiology	10	50	40	10	33-	57	15	34	15	10	10	44	19
Basic-, supportive subjects	**59**	**58**	**36+**	**7-**	**44**	**58**	**9**	**29**	**7**	**13**	**8**	**47**	**10**
Epidemiology/statistics	7	71+	29	0-	57+	55	23	32	15	12	21	41	26
Metamedica	5	20--	80++	0-	-23--	51	23	38	23	11	14	32	33
Psychiatry/psychology	12	67	33++	0-	54-	73	14	20	12	7	10	65	17
Social medicine	3	33	67+	0-	0	42	26	51	28	8	15	17	37
Behavioural scientific/other subjects	**27**	**56**	**44++**	**0-**	**35-**	**61**	**13**	**30**	**9**	**9**	**10**	**47**	**15**
Surgery	16	69	31	0-	56	67	13	27	12	6	8	56	17
Dermatology/ENT/ opthalmology	14	57	36	7	44	63	14	29	14	8	10	53	18
Geriatrics	8	62	38+	0	44	68	17	29	16	3	6	55	23
Obstetrics/Gynaecology	7	43--	57++	0-	21--	60	14	28	17	13	14	49	19
Family medicine	20	40--	55++	5	21--	61	12	34	12	4	5	49	16
Internal medicine	26	73	19	8+	67	73	11	22	9	5	5	64	14
Paediatrics	12	50-	42++	8	32-	60	15	28	13	12	12	48	19
Neurology	7	43	43+	14	19-	50	17	32	17	18	19	37	21
Clinical subjects	110	57-	37++	5	43--	65	8	28	7	7	6	54	10

-/--/++/+ low respectively high in comparison with the total group. Results are presented in percentages. Std = standard deviation. ? = question mark option use.

Appendix 2 – Questionnaire

In the context of the study on the effect of different assessment conditions of the progress test on learning behaviour we would like to conduct a short questionnaire about the last progress test on the 2nd of February 2022.

This questionnaire **only** concerns the **progress test on the 2nd of February 2022.**
ProF refers to the online feedback system of the progress test.

For part of the students the result of this progress test did not count towards the awarding of credits, while for another part of the students it did. Indicate what applies to you.

1. Did the result of this progress test count towards the awarding of credits?
    - o Yes
    - o No
    - o Don't know

2. How important was this progress test for you (e.g. for obtaining credits, for your study progress, personal reasons). Choose one answer option.
    - o Low
    - o Intermediate
    - o High

The following questions relate to the preparation prior to the progress test on the 2nd of February 2022.

3. Did you prepare for this progress test? Choose one answer option.
    - o Yes
    - o No

*In case you answered **question 3** with "**yes**", you can continue with question 5 and skip question 4. In case you answered **question 3** with "**no**", continue with question 4.*

4. Why did you not prepare for this VGT? Multiple answers possible.
    - o I had no time to prepare
    - o I did not feel like preparing
    - o I always pass my progress test without preparation
    - o I got a pass/good for my previous progress test
    - o I thought this progress test was not important
    - o Other: .....

The following questions relate to the consultation of the feedback after the progress test on the 2nd of February 2022.

5. Did you check the answers of this progress test with the answer key?
    - o Yes
    - o No

6. Did you look at the feedback of this progress test in the email?
   - o Yes
   - o No

7. Did you consult ProF to look at the feedback of this progress test?
   - o Yes
   - o No

*In case your answer to **question 9** was "**no**", you can continue with question 11 and skip question 10. In case you answered "**yes**" to **question 9**, you can continue with question 10 and skip question 11.*

8. What is the reason that you did not look at the feedback in ProF? Multiple answers possible.
   - o I do not know where I can find the feedback
   - o I did not have time to look at the feedback
   - o I did not put effort in this progress test
   - o I thought this progress test was not important
   - o I had a pass/good for this progress test
   - o I find the feedback not useful
   - o I already checked my answers with the answer key
   - o Other: .....

9. Which section of ProF did you look at? Multiple answers possible.
   - o Progress total score (longitudinal)
   - o Total score of this progress test (moment)
   - o Progress on discipline score (longitudinal)
   - o Discipline score of this progress test (moment)
   - o Progress on category score (longitudinal)
   - o Category score of this progress test (moment)
   - o Progress on cluster score (longitudinal)
   - o Cluster score of this progress test (moment)
   - o I do not know

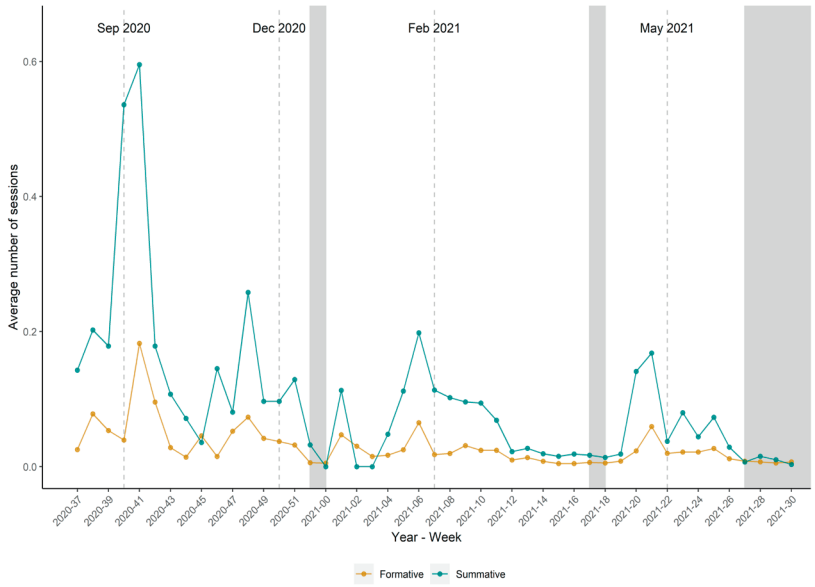
*The following seven items relate to the **consultation of feedback in the email or ProF after** the progress test on the 2nd of February 2022. For the statements below, indicate the extent to which you agree or disagree with each statement (1 = strongly disagree, 2 = disagree, 3 = somewhat disagree, 4 = either agree or disagree, 5 = somewhat agree, 6 = agree, 7 = strongly agree).*

1. I actively use the feedback to help me improve.
2. I pay attention to the feedback.
3. I use the feedback to set goals for the next progress test.
4. I look at the feedback to see what I did wrong.
5. The feedback makes me try harder.
6. The feedback changes the way I learn and study.
7. I enjoy getting the feedback.

Appendix 3 – Interview guide

Part 1. Own feedback experiences
1. Do you prepare for the progress test?
How do you prepare?
What determines whether you prepare for the progress test?
2. Do you consult the result of the progress test?
Which methods do you use to consult the test result?
What determines whether you look at the test result? And what is the role of the test condition in this?
3. Do you use the result of the progress test?
What do you do with this information?
What determines whether you use the feedback? And what is the role of the test condition in this?
4. Are you aware of the online feedback system (ProF)? Only asked if students did not mention ProF yet.
Why are you not using this feedback system?
What do you think is the reason that you are not aware of ProF?
Part 2. Reflection and interpretation of ProF logging data (graph)
1. Can you describe what you see?
2. What do you think when you see these data?
3. How would you explain and/or interpret these data?
Part 3. Perception of progress test and feedback
1. What is your perception of the progress test? And which place does it have in your study program?
2. What is your perception of the way(s) the test result is presented to you?
Do you have any suggestions to improve this?

Appendix 4 – Supplemental Figure 1



Supplemental Figure 1. The average number of ProF sessions for the progress tests of September 2020, December 2020, February 2021, and May 2021 divided in students who participated in the formative (yellow line) or summative condition (blue line) which is shown to the interviewees for reflection and possible explanations for the trend.

Appendix 5 – Supplemental Table 2

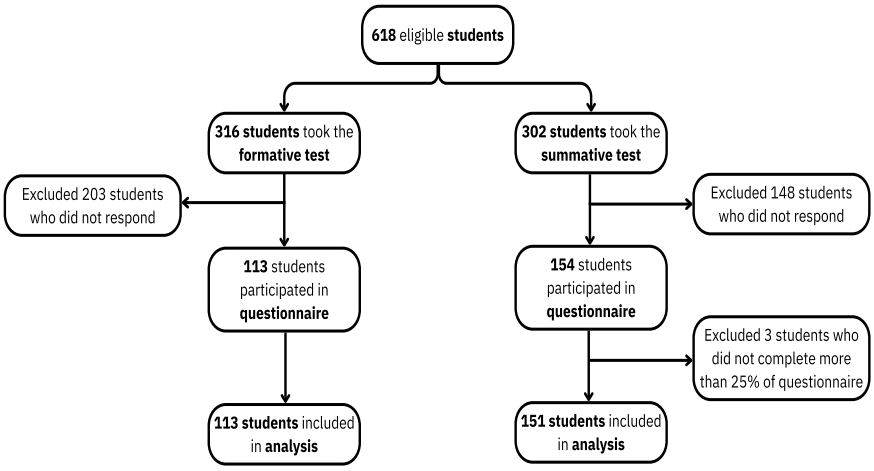
Supplemental Table 2. Descriptive characteristics of interviewees.

Number of logging sessions[a]	0		1		2 to 4		>5		
Grade	Fail	Pass/ Good	Fail	Pass/ Good	Fail	Pass/ Good	Fail	Pass/ Good	Total year (M/F)
Year 2	11[c]	3[c], 10[c]		21[c]		20[b],19[c]			6 (1/5)
Year 3	7[c]	18[c]	4[b]	9[c]		5[b], 8[b]	2[c]	15[b]	8 (4/4)
Year 5	6[c]		14[c]		16[c]	17[b]			4 (1/3)
Year 6	13[b]			1[c]		12[b]			3 (2/1)
Total Fail and Pass/Good (M/F)	4 (1/3)	3 (0/3)	2 (1/1)	3 (0/3)	1 (0/1)	6 (5/1)	1 (0/1)	1 (1/0)	

[a]ProF logging sessions from September 2020 to January 2021. M = male, F = female.

[b]Male. [c]Female.

Appendix 6 – Supplemental Figure 2



Supplemental Figure 2. Flowchart of the participants of the questionnaire.

Appendix 7 – Supplemental Table 3

Supplemental Table 3. Baseline characteristics of responders and non-responders of the questionnaire.

	Responders (*n*=264)	Non-responders (*n*=354)	SMD[a]
Age, median (IQR)	21 (20, 21)	21 (20, 22)	0.107
Female, n (%)	185 (70)	244 (69)	0.022
Grade, n (%)			
Fail	24 (9)	64 (18)	0.266
Pass	106 (40)	196 (55)	0.304
Good	134 (51)	94 (27)	0.508
Proportion passed earlier PTs, %			
September 2021	223 (87)	286 (85)	0.058
December 2021	226 (87)	249 (73)	0.355
ProF sessions, mean (SD)[b]	1.29 (1.60)	0.71 (1.87)	0.334

SMD, standardized mean difference. IQR, interquartile range. PT: progress test.

[a]A standardized mean difference >0.1 may point towards meaningful imbalance between groups.

[b]Period of consultation for each PT starts the week before the PT and ends the week before the next PT.

Appendix 8 – Supplemental Table 4

Supplemental Table 4. Students' idea of assessment condition and perceived stakes of the progress test.

	Formative Test (*n*=113)	Summative Test (*n*=151)
Students' idea of condition (Q1), n (%)		
Formative	79 (70)	8 (5)
Summative	11 (10)	128 (85)
Don't know	23 (20)	14 (9)
Perceived stakes (Q2), n (%)		
Low	57 (50)	19 (13)
Intermediate	48 (42)	94 (62)
High	8 (7)	37 (25)

Q1: *Did the result of this progress test count towards the awarding of credits?*

Q2: *How important was this progress test for you (e.g. for receiving credits, for your study progress, personal reasons)?*

Appendix 9 – Supplemental Table 5

Supplemental Table 5. Reasons for not preparing for the *formative* or *summative* progress test.

	Formative Test	Summative Test	p-value[a]
True formative and summative			
Number of individuals	99	123	
No preparation, n (%)			
Time[c]	35 (35)	42 (34)	0.765
Motivation	18 (18)	11 (9)	0.036
Need	69 (70)	92 (75)	0.543
Grade	39 (39)	46 (37)	0.671
Importance	27 (27)	1 (1)	<0.001
Other	5 (5)	6 (5)	1.000[d]
Perceived formative and summative[b]			
Number of individuals	71	103	
No preparation, n (%)			
Time	27 (38)	35 (34)	0.491
Motivation	14 (20)	10 (10)	0.050
Need	45 (63)	78 (76)	0.134
Grade	29 (41)	39 (38)	0.584
Importance	26 (37)	0 (0)	<0.001
Other	4 (6)	4 (4)	0.715[d]

[a]Chi-squared test.

[b]Subgroup analysis; Perceived formative = students in the formative test group who knew it was formative; Perceived summative: students in the summative test group who knew it was summative.

[c]Time: *"I had no time to prepare"*; Motivation: *"I did not feel like preparing"*; Grade: *"I got a pass/good for my previous progress test"*; Importance: *"I thought this progress test was not important"*.

[d]Fisher's exact test.

Appendix 10 – Supplemental Table 6

Supplemental Table 6. Average ProF logging sessions for each progress test

	Number of individuals	Average ProF sessions, mean (95% CI)	p-value[b]
PT September 2021[c]			
Year 2	316	0.67 (0.52-0.82)	
Year 3	305	1.25 (1.03-1.46)	
PT December 2021			
Year 2	316	0.38 (0.19-0.57)	
Year 3	305	0.61 (0.44-0.79)	
PT February 2022[d]			
Year 2	316	0.39 (0.27-0.51)	
Year 3	305	0.83 (0.66-1.01)	<0.001

ProF = progress test feedback system; PT = progress test.

[a]Population based on participants of the PT in February.

[b]Unpaired t-test.

[c]Period of consultation for each PT starts the week before the PT and ends the week before the next PT.

[d]PT February 22 summative for year 3, formative for year 2.

Appendix 11 – Supplemental Table 7

Supplemental Table 7. Median scores (IQR) of the 6-point Likert scale items of *Active use of feedback* and its subfactor *Enjoyment* in the *formative* and *summative* progress test-group.

	Formative Test	Summative Test
True formative and summative		
Number of individuals[a]	91	135
Feedback use, median (IQR)		
Item 1[c]	3 (2-3)	3 (2-3)
Item 2	3 (3-4)	3 (2-4)
Item 3	2 (1-3)	2 (1-3)
Item 4	4 (3-5)	4 (3-5)
Item 5	3 (2-4)	3 (2-4)
Item 6	2 (1-3)	2 (1-3)
Item 7	5 (4-6)	5 (4-6)
Perceived formative and summative[b]		
Number of individuals	62	114
Feedback use, median (IQR)		
Item 1	3 (2-3)	3 (2-3)
Item 2	3 (3-4)	3 (2-4)
Item 3	2 (1-3)	2 (3-4)
Item 4	4 (3-5)	4 (3-4)
Item 5	3 (2-4)	3 (2-4)
Item 6	2 (1-3)	2 (1-3)
Item 7	5 (4-6)	5 (4-6)

IQR = interquartile range.

[a]Students who consulted feedback in e-mail or progress test feedback system.

[b]Subgroup analysis; Perceived formative= students in the formative test group who knew it was formative; Perceived summative = students in the summative test group who knew it was summative.

[c]Item 1 = I actively use the feedback to help me improve; Item 2 = I pay attention to the feedback; Item 3 = I use the feedback to set goals for the next progress test; Item 4 = I look at the feedback to see what I did wrong; Item 5 = The feedback makes me try harder; Item 6 = The feedback changes the way I learn and study; Item 7 = I enjoy getting the feedback.

Part III: Feedback

Chapter 9

Understanding students' feedback use in medical progress testing: A qualitative interview study

Elise V. van Wijk
Floris M. van Blankenstein
Roemer J. Janse
Eline A. Dubois
Alexandra M.J. Langers

Medical Education. 2024;58(8):980-988.
DOI: 10.1111/medu.15378.

Abstract

Background Active engagement with feedback is crucial for feedback to be effective and improve students' learning and achievement. Medical students are provided feedback on their development in the progress test (PT), which has been implemented in various medical curricula, although its format, integration and feedback differ across institutions. Existing research on engagement with feedback in the context of PT is not sufficient to make a definitive judgement on what works and which barriers exist. Therefore, we conducted an interview study to explore students' feedback use in medical progress testing.

Methods All Dutch medical students participate in a national, curriculum-independent PT four times a year. This mandatory test, composed of multiple-choice questions, provides students with written feedback on their scores. Furthermore, an answer key is available to review their answers. Semi-structured interviews were conducted with 21 preclinical and clinical medical students who participated in the PT. Template analysis was performed on the qualitative data using a priori themes based on previous research on feedback use.

Results Template analysis revealed that students faced challenges in crucial internal psychological processes that impact feedback use, including 'awareness', 'cognizance', 'agency' and 'volition'. Factors such as stakes, available time, feedback timing and feedback presentation contributed to these difficulties, ultimately hindering feedback use. Notably, feedback engagement was higher during clinical rotations, and students were interested in the feedback when seeking insights into their performance level and career perspectives.

Conclusion Our study enhanced the understanding of students' feedback utilization in medical progress testing by identifying key processes and factors that impact feedback use. By recognising and addressing barriers in feedback use, we can improve both student and teacher feedback literacy, thereby transforming the PT into a more valuable learning tool.

Introduction

Effective feedback improves students' learning and achievement, fosters adaptive learning and prepares students for life-long learning [1-3]. The longitudinal medical progress test (PT) is implemented in various countries around the world [4]. Although PTs are operationalised in different ways across medical schools, their core function is to provide medical students feedback on their knowledge growth throughout their studies. This feedback is aimed at assessment *for* learning as it stimulates identification of learning gaps and adjustment of learning [1, 5-7]. Feedback is effective when students actively engage with it and act upon it. However, in practice, this rarely happens, which raises doubts about the effectiveness of PT feedback [8-13].

To promote student engagement with feedback, students need to acquire feedback literacy, that is, be able to understand, appreciate, utilise and benefit from feedback processes. This requires a proactive attitude and a shift towards a learning-centred approach [14-16]. Based on the framework of Carless and Boud, [14] Molloy *et al.* [15] identified seven essential categories of student feedback literacy, including the understanding of feedback purposes, emotional engagement, and utilising the feedback for future work. Furthermore, Winstone *et al.* [17] identified four pivotal internal psychological processes for effective feedback use: (1) *awareness* of what feedback means and what its purpose is, (2) *cognizance* of appropriate strategies to implement feedback, (3) *agency* in implementing these strategies and (4) *volition* (or will) to explore and act on the feedback. Difficulties in these processes form barriers to effective feedback use. In undergraduate medical education, feedback receptiveness is primarily influenced by students' characteristics (e.g. confidence and mindset), feedback content, educators' credibility and the learning environment [18]. Guidelines for effective feedback in clinical learning underscore the impact of clear and supportive feedback on motivating trainees. The most important elements include enhancing self-efficacy and fostering the development of strategies that lead to improved competencies [19]. Additionally, teachers and institutions play a crucial role in fostering a climate that supports learning and enhances engagement with feedback [16, 18-21].

Unlike most assessments, the PT has a repetitive, comprehensive and curriculum-independent nature. The feedback of the PT has an important formative function, aiming to guide and improve students' learning. However, the effectiveness of PT feedback remains uncertain [8, 10-13]. Students often struggle to comprehend or utilise PT feedback [11], and they may lack the agency to translate the feedback into actionable strategies [12]. Agency and its importance in feedback use was also highlighted by Winstone *et al.* [17]. When both PTs and end-of-course tests have summative purposes, students tend to focus on the latter and be less inclined to self-regulate their learning with PT feedback [8]. Furthermore, research shows that students rarely use the PT to reflect and improve their learning [13]. These studies indicate a significant loss of the formative value that PT feedback is intended to provide. Still, acceptance of progress testing is enhanced by sufficient, detailed, personalised, well-timed and specific feedback [4]. However, current literature lacks studies about students' actual use of PT feedback and the factors influencing this feedback use. Therefore, this qualitative study aimed to explore which processes and factors affect medical students' feedback use within a Dutch progress testing context. By exploring this, actions can be undertaken to overcome obstacles, support students' feedback use and improve student feedback literacy.

Methods

Setting

This study was set in a 6-year medicine programme (split into a 3-year preclinical bachelor and 3-year clinical master) at Leiden University Medical Centre (LUMC). In the Netherlands, all medical students are required to participate in a national, curriculum-independent PT four times annually, in addition to their regular course assessments. This results in a total of 24 PT test moments over the course of their study. The PT is a written test consisting of multiple-choice questions (MCQs) that cover all relevant medical disciplines and are stratified into categories. The MCQs include a question mark option that yields no points. Points are deducted if an answer is incorrect [22]. Students receive a *'Good'*, *'Pass'* or a *'Fail'* on each PT, and at the end of each academic year, the scores are combined and translated into a summative decision (i.e. study credits). After each PT, the answers are available online, and students receive feedback via e-mail. Students can also access feedback in the online Progress test Feedback System (ProF) [23]. The feedback is presented as an individual score stratified by category and discipline compared with their peer students. This is displayed in the e-mail (*Appendix 1 – Supplemental Table 1*) and graphically in ProF [23]. Students can reflect on the feedback voluntarily with their tutor, who is a medical specialist who offers support throughout the programme. Normally, all students take the PT in a lecture hall under live supervision. During the COVID-19 pandemic, when this study took place, some of the PTs were taken from home via a digital assessment platform. Due to logistic reasons, not all online PTs could be proctored by the online proctoring software. These non-proctored PTs were turned into formative PTs (e.g. no impact on obtaining study credits) because students were able to consult study materials for answering the questions. The online proctored PTs remained summative, just as the PTs taken in the lecture hall.

Study design

Based on the subtle realism paradigm, a qualitative study design was chosen. This paradigm combines a realist ontology with a constructivist epistemology, acknowledging an objective reality independent of our perceptions, while also recognising that our understanding of this reality is shaped by our subjective perspectives [24, 25]. The truth is negotiated through dialogue and the goal of 'objectivity' is an ideal for which to strive. Rigorous research methods, such as purposive sampling, member checking and reflexivity are used to enhance the objectivity and credibility of research findings. Subtle realism also aims to provide a logical and coherent interpretation of the data (i.e. plausibility), instead of proving a definitive cause-and-effect relationship. This paradigm allowed us to provide a nuanced and comprehensive exploration of students' feedback use. We developed an interview guide with open-ended questions about students' feedback behaviour in relation to the PT (*Appendix 2 – Interviewguide*). This study was part of a more comprehensive interview study, in which we also assessed the effects of PT assessment conditions (formative versus summative) on feedback behaviour. After two pilot interviews with fourth-year medical students, individual semi-structured interviews were conducted in April and May 2022.

Sampling and data collection

Medical students were included in the sampling if they participated in at least four out of six PTs in September 2020 to December 2021 and in both formative and summative PT assessment conditions. Maximum variation sampling based on the frequency of ProF use, PT results and study year was used to ensure the representation of multiple perspectives [26]. We first made proportional groups based on

the distribution of ProF logging sessions among all students who participated in the PTs. Within these groups, we aimed to sample an equal number of students from different study years with a *'fail'* or *'pass/good'*. If a student had failed at least one PT, the student was assigned to the 'fail' group (n=410). The other students were assigned to the *'pass/good'* group (n=876). All student data were derived from the university's student administration system. Participants were invited by e-mail and received an electronic gift card in return for participation. EvW conducted 21 interviews of 30–60 min via online meetings in Microsoft Teams.

Data analysis

The audiotaped interviews were transcribed verbatim and pseudonymized before analysis. Template analysis was used to analyse the interviews, initially after four interviews (when small adjustments were made to the interview guide) and then after all the interviews were completed. Template analysis uses existing literature to formulate a priori themes that guide deductive analysis and the development of successive coding templates [27]. Our a priori themes (*Table 1*) were based on the psychological processes underlying barriers in feedback use as described by Winstone *et al.* [17] and other literature on feedback use [1, 28, 29]. Two independent coders (EvW and FvB) coded Interviews 1–6 in ATLAS.ti 22.0.11.0 for Windows [30]. Afterwards, EvW and FvB discussed the obtained codes together with a third researcher (AL), who coordinates the PT within the LUMC and could therefore provide feedback on the validity and clarity of the initial template from her own expertise. EvW and FvB then coded Interviews 7–14, revised the initial template again and coded Interviews 15–21, after which they determined the final template. This procedure yielded only minor revisions in the initial template. No new themes emerged in the final template, indicating theoretical sufficiency after Interview 14 [31, 32]. The final template (*Table 2*) was discussed with the research team (EvW, FvB and AL). Eventually, EvW reread and recoded all interviews with the final template to ensure all relevant information to answer the research question was included in the template. In this final phase, it became evident that most of the codes did not exist in isolation but co-occurred frequently. EvW coded the relations between these co-occurring codes, discussed these new relations with FvB and leveraged insights from these discussions to develop a thematic map (*Figure 1*). Synthesised Member Checking (SMC) method, which is suggested as an appropriate method within a subtle realism paradigm [33], was performed and yielded the additional theme *'Guidance'*.

Table 1. A priori themes used in the template analysis.

	***A priori* themes**
1.	Awareness
2.	Cognizance
3.	Agency
4.	Volition
5.	Score
6.	Stakes
7.	Time
8.	Utility value

Table 2. Final template.

	External factors
1.	**Stakes**
	Test moment
	Grade
2.	**Time available**
3.	**Feedback timing**
4.	**Feedback message**
	Presentation
	Specificity
5.	**Guidance**
	Internal processes
1.	**Awareness**
	Inability to decode feedback
2.	**Cognizance**
	Poor knowledge of appropriate opportunities
	Poor knowledge of appropriate strategies
3.	**Agency**
	Sense of disempowerment
	Difficulties with translating feedback into action
4.	**Volition**
	Lack of interest
	Lack of or change in priority
	Clinical phase
	Other (study) activities
	Grade focus
	Enhancement of self-knowledge
	Performance level
	Future career perspectives

Reflexivity

Researcher reflexivity is required to account for how subjective perspectives shaped the qualitative research process [34]. Our reflexivity statement can be found in *Appendix 3*.

Results

Participant demographics are shown in *Appendix 4 – Supplemental Table 2*. The thematic map illustrates the relations between our themes and codes that describe the processes and factors involved in PT feedback use (*Figure 1*). The findings will be presented as a comprehensive narrative, structured by Winstone *et al.*'s internal psychological processes [17].

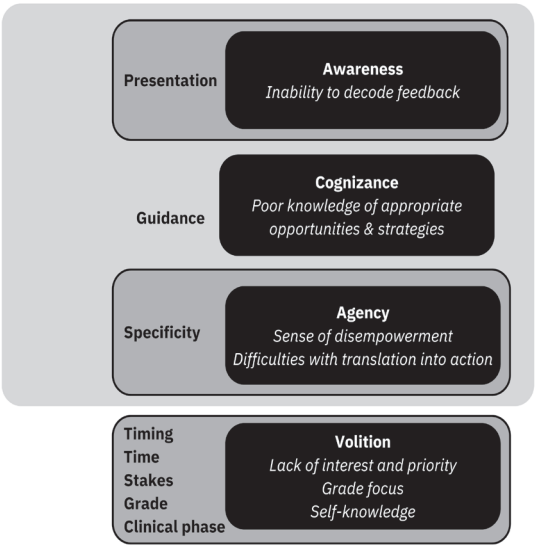


Figure 1. The internal processes (bold white text) with their underlying barriers (italic white text) to feedback recipience are represented as black boxes. The factors that affect the feedback engagement are displayed in the dark grey boxes. Guidance (light grey box) is suggested by students as means to address the challenges related to *awareness, cognizance,* and *agency.*

Awareness

Awareness can be conceived as knowing what feedback means and what its purpose is [17]. Almost all students expressed frustration regarding the complexity of the feedback in the e-mail (*Appendix 1 & 2 – Supplemental Table 1 & 2*), which was mainly caused by the unclear ***presentation.*** As they were ***not able to decode this feedback***, the feedback was not used:

"I really never understood that overview in the e-mail. I don't use it, because I think it is such an abracadabra with all those pluses and minuses." (Interview #12, Y6)

The majority of students indicated that the feedback in ProF, which is visually presented (also see *Supplemental Figures 1 and 2*), was clearer and therefore easier to understand:

"I find ProF very useful, because you can easily see at a glance in a graph how well you have scored." (Interview #8, Y3)

Cognizance

Cognizance means that students know about opportunities for seeking support in using feedback and beneficial learning strategies [17]. Several students did not realise they could discuss the feedback with their tutor or that they could access the feedback in ProF. This ***poor knowledge of opportunities*** prevented them from using the feedback.

However, even students who knew about ProF admitted that they often did not take advantage of it as they were ***unaware of which strategies were beneficial*** to successfully implement the feedback. Many students lacked knowledge on how to use received feedback:

"My consideration to use the feedback, or prepare for the next PT, depended on how easy it was to study, how broad the topic was. Because, if it's, for example, you've made too many mistakes in internal medicine. Yes, that (topic) is so incredibly broad. I just cannot study that." (Interview #13, Y6)

Agency

Even when students are cognizant, they still require confidence in the implementation, which is defined as *agency* [17]. Most students failed to ***translate the feedback*** by e-mail ***into action*** because of a lack of understanding. Even for students who did understand, or who consulted ProF, the feedback was considered not specific enough making it unclear where to start and how to use it for further learning:

"It (the feedback) is very broad, and there are more questions for each category, and you actually do not know what to do, well there are big differences between the subjects of the questions." (Interview #1, Y6)

"Yeah, I kind of don't know where to start. For example psychiatry, I have not scored very well on that lately. Then I think: 'I have no clue of what I can do. Should I grab the Diagnostic and Statistical Manual of Mental Disorders, or something?' It would help if we get clear instructions on how to practice, and use the feedback." (Interview #17, Y5)

Furthermore, many students value the potential of feedback to improve themselves, which gives them a sense of empowerment [17]. However, most students felt ***disempowered***, which hindered them from using the feedback, especially when they already passed the PT. This feeling was related to their focus on subject content [35], and direct transferability of knowledge to the next PT rather than appreciating the more long-term development on the different knowledge domains for which the PT is intended.

Volition

To act on feedback, students need to be '*ready to engage*', [36] and thereby show volition: willingness to perform action [17]. Several students were not willing to invest energy in the feedback, because they lacked interest, or gave no priority to the PT. Their perceived late ***timing*** of the feedback (2 weeks after the PT) reinforced their lack of interest, which suggests that they would rather use the feedback to directly evaluate their performance (score) than to monitor their own growth and learning process. Almost all students mentioned that they had too little ***time available*** in their (study) schedule to pay attention to the feedback. They usually spent their time on end-of-course assessments, or social activities, which they prioritised over consulting the PT feedback:

"Because we are busy and also have other things, such as examinations, then you think, if this goes well without repetition then yes, you do not want to spend extra time on the feedback." (Interview #18, Y3)

PT feedback gained priority when the ***stakes*** (i.e. consequences for study progress) were higher. Stakes were affected by passing, or failing a PT, and by the test moment.

For instance, the last PT of the bachelor or master (Test moment 12 or 24) was high stakes for most students, which motivated students to use previous PT feedback to prepare for the next PT:

"I think that (test moment 24) also played a role, because I think that people are stressed about doing well so they can finish their study. And that is why they also use ProF to look at earlier feedback and be able to prepare well." (Interview #7, Y3)

Also, the ***grade*** was an important determinant in their volition to act on the feedback. Through this *'grade focus'* [17], an insufficient grade incentivised acting on the feedback, whereas a sufficient grade incentivised consulting the feedback superficially, or ignoring it.

Some students changed their attitude towards their study after the transition to the ***clinical phase***, which was accompanied by an increasing interest in the feedback:

"I think because I took my studies more seriously in the master compared to the bachelor. In the bachelor, it was more like: 'Pass, fine, check, and move on.' In the master, it was more like: 'Okay, what can I still learn from my mistakes?'" (Interview #20, Y2)

Students who were curious about the feedback mainly used it to enhance their self-knowledge, confirm their overall performance or check their performance on domains of future career interests:

"You're almost at the end of your bachelor's degree, and you're looking at which specialty to choose. At this moment, I really like endocrinology and gynecology, so when I look at the different domains on ProF, I will first have a closer look at the scores on those domains, and ask myself 'Is it justified that I like those specialties?'"(Interview #2, Y3)

Guidance

A lack of guidance limited many students in their feedback use. Students mentioned that they would appreciate more opportunities to practice with PT questions in the curriculum, although there is an existing possibility to practice previous PT questions on the national PT website. They preferred a more detailed explanation of the current possibilities provided by ProF, what the feedback means and how they can use it:

"You could make some practice questions available, so you can see the connection between the PT and the courses and the PT is more integrated in the curriculum. And if, for example, you do not know something, you can directly explore this within the course. I still miss that a bit." (Interview #5, Y3)

Discussion

This study provides an overview of relevant processes and factors that can prevent or stimulate feedback use in medical progress testing and how these processes and factors relate to each other (*Figure 1*). Most students did not understand the feedback, had poor knowledge of appropriate opportunities and strategies to use the feedback, felt disempowered and insecure about translating the feedback into action, lacked interest and tended to focus on grades.

In contrast to Winstone *et al.* [17], we did not find evidence that students were limited in their feedback use by narrow conceptions of the feedback's purpose. Students seemed well aware that the feedback was meant for self-reflection and could help them grow, in line with the assessment for learning. Moreover, students were willing to consult feedback to learn about their performance in relation to future career perspectives. Students in the clinical phase were also more engaged with the feedback.

Students experienced difficulties in understanding the PT feedback (*awareness*) [15, 17], mainly because the presentation of the feedback in the e-mail was unclear. The visual presentation of the feedback in ProF improves the level of understanding and thereby feedback engagement. In general, students had poor knowledge of the opportunities, such as reflection with the tutor, and strategies to utilise the feedback (*cognizance*) [17], which hindered feedback use. Difficulties with translating the feedback into action (*agency*) [17] were attributed to the lack of specificity of the feedback message, which aligns with prior findings in the PT [11, 12]. In undergraduate medical education, negative perceptions of students also relate to unspecific and unclear feedback [18]. Students reported that they needed more guidance and support to overcome these barriers; they preferred clear and explicit instructions and tools on how to read and handle the feedback. This struggle with appropriate learning strategies and the desire for explicit instructions is also reported in earlier studies [17, 37-39]. Guidance or explicit suggestions on how to address knowledge gaps could enable students to construct appropriate strategies for improvement [19].

Most students were not willing to act upon the feedback (*volition*) [17], primarily due to its perceived insignificance and lack of interest. The delayed timing of the feedback contributed to this lack of interest, indicating a disregard for its importance in enhancing the long-term learning process. The immediate availability of the answer key after the PT was appreciated by the students. Remarkably, several students mistakenly believed that the answers were only available for 24 hours (which actually was the time to appeal for mistakes in the questions and/or answer key), which hindered its use. Several studies show that feedback timing is important, with immediate feedback generally leading to higher task satisfaction [1, 4, 40, 41]. However, most students did not prioritise the PT feedback, and instead, focused on the end-of-course assessments. This aligns with earlier findings where end-of-course assessments were perceived as being higher stakes and more rewarding than the PT [8]. This preference may be linked to the PT's position as a curriculum-independent test and its lack of integration in the curriculum, which is known to be important in promoting student feedback engagement [15, 16, 18-20, 42]. Conversely, higher stakes test moments and insufficient grades on the PT led to increased feedback utilisation. This contrasts with findings of more feedback engagement in high-performing students after objective structured clinical examinations (OSCEs) [43], yet aligns with the concept of *'grade focus'* [17, 44, 45], which has been extensively discussed in the literature as a factor affecting students' receptiveness to and engagement with feedback. Furthermore, the pass/fail decision associated with the PT may contribute to its perception as a purely summative assessment, potentially impeding the intended formative function of the feedback [16, 46].

Feedback engagement was enhanced in students in the clinical phase and in students willing to gain knowledge on their performance and future career perspectives. Although the self-assessment was mainly limited to feedback consultation (a more passive feedback use), this finding provides insights into student incentives to use PT feedback.

Our study demonstrates that the pre-clinical to clinical transition is accompanied by a more serious attitude towards the study and an increased interest in the PT feedback. The feedback can be extra valuable in the clinical phase to address knowledge gaps and apply theoretical knowledge in practice [47, 48].

Strengths and limitations

We positioned our study in the existing literature on feedback use by using a previously established theoretical framework [17] and expanded this framework to elucidate all relevant factors that play a role in the use of PT feedback use. Our sample encompassed a representative sample of medical students from both preclinical and clinical stages with a variety in ProF utilization and PT results. Our findings align with concepts found in earlier feedback research in different educational contexts, encompassing key aspects such as the internal psychological processes [17], specificity [18, 19], feedback timing [1, 4, 40, 41] and grade focus [17, 44, 45]. This convergence suggests that the present findings may also be valid in other settings in which assessment *of* learning is combined with assessment *for* learning, such as a programmatic assessment setting where students are encouraged to use feedback from low-stakes tests to enhance their future learning. Nevertheless, it is crucial to acknowledge that unique features of the PT, such as its' repetitive and comprehensive nature, may affect how students engage with feedback. Additionally, different operationalizations of the PT may affect feedback use as well, as it is shown to influence learner acceptance [4]. For example, other medical schools may put more emphasis on guidance through tutoring, which is shown to stimulate feedback use [42]. Finally, we did not address issues related to Equity, Diversity, and Inclusion (EDI), as these (participant) data were not registered, and therefore not available. Future research could benefit from exploring these issues, as students with specific learning difficulties and from minority groups might face particular challenges with written and graphical feedback.

Implications for practice

The present findings can be used to address and overcome barriers hindering the use of PT feedback, thereby enhancing students' feedback literacy and effective feedback use. These barriers can be overcome by timely feedback provision, clear and specific feedback presentation and guidance to aid students in its effective use. Utilising feedback prompts that describe the learning objectives of the questions could enhance the specificity of the feedback, as shown by Burr *et al.* [49] Even though self-regulation is becoming more important in medical education [50, 51], our study reveals that most students do not use the opportunity to reflect on the feedback with their tutor and may still require explicit instructions and guidance in feedback use. This guidance can be offered by teachers and/or the institution early in the curriculum, as it is known that this shared responsibility is crucial in developing student feedback literacy skills [15, 52]. This may also require a better integration of the PT into the curriculum, as suggested by the interviewed students and earlier studies [42, 46, 53]. Interactive dialogue between students and teachers is important here instead of static information provision [21]. This aligns with key principles of programmatic assessment, in which there is a strong focus on feedback literacy and dialogue. Under such circumstances, students might be more receptive to engage with PT feedback [16, 42].

Conclusion

This study demonstrates that PT feedback use by medical students is hampered by the experience of difficulties related to awareness, cognizance, agency and volition. The stakes, available time, feedback timing, feedback message and lack of guidance contribute to these difficulties and can further prevent feedback use. Student feedback engagement could be enhanced by providing guidance and explicating its relevance for self-regulated learning. Additionally, effective communication and integration of the PT and its feedback within the curriculum can further promote feedback engagement, elevating the PT's significance as a valuable learning instrument.

References

1. Hattie J, Timperley H. The Power of Feedback. Review of Educational Research. 2007;77(1):81-112.
2. Koh LC. Refocusing formative feedback to enhance learning in pre-registration nurse education. Nurse Education in Practice. 2008;8(4):223-30.
3. 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.
4. Dion V, St-Onge C, Bartman I, Touchie C, Pugh D. Written-Based Progress Testing: A Scoping Review. Academic Medicine. 2022;97(5):747.
5. Nicol DJ, Macfarlane-Dick D. Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. Studies in Higher Education. 2006;31(2):199-218.
6. Black P, Wiliam D. Assessment and Classroom Learning. Assessment in Education: Principles, Policy & Practice. 1998;5(1):7-74.
7. Schuwirth LWT, Van der Vleuten CPM. Programmatic assessment: From assessment of learning to assessment for learning. Medical Teacher. 2011;33(6):478-85.
8. Van Berkel HJM, Nuy HJP, Geerligs T. The influence of progress tests and block tests on study behaviour. Instructional Science. 1994;22(4):317-33.
9. 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.
10. Aarts R, Steidel k, Manuel BAF, Driessen EW. Progress testing in resource-poor countries: A case from Mozambique. Medical Teacher. 2010;32(6):461-3.
11. 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.
12. 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.
13. 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):Doc41.
14. Carless D, Boud D. The development of student feedback literacy: enabling uptake of feedback. Assessment & Evaluation in Higher Education. 2018;43(8):1315-25.
15. Molloy E, Boud D, Henderson M. Developing a learning-centred framework for feedback literacy. Assessment & Evaluation in Higher Education. 2020;45(4):527-40.
16. Watling CJ, Ginsburg S. Assessment, feedback and the alchemy of learning. Medical Education. 2019;53(1):76-85.
17. 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.
18. Cordovani L, Tran C, Wong A, Jack SM, Monteiro S. Undergraduate Learners' Receptiveness to Feedback in Medical Schools: A Scoping Review. Medical Science Educator. 2023;33(5):1253-69.
19. Lefroy J, Watling C, Teunissen PW, Brand P. Guidelines: the do's, don'ts and don't knows of feedback for clinical education. Perspectives on Medical Education. 2015;4(6):284-99.
20. Ramani S, Könings KD, Ginsburg S, van der Vleuten CPM. Meaningful feedback through a sociocultural lens. Medical Teacher. 2019;41(12):1342-52.
21. Ajjawi R, Regehr G. When I say … feedback. Medical Education. 2019;53(7):652-4.
22. Lord FM. Formula scoring and number-right scoring. Journal of Educational Measurement. 1975;12(1):7-11.
23. 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.
24. Maxwell JA, Mittapalli K. Realism as a Stance for Mixed Methods Research. SAGE Handbook of Mixed Methods in Social & Behavioral Research: SAGE Publications, Inc.; 2010. p. 145-68.
25. Fetters MD, Curry LA, Creswell JW. Achieving Integration in Mixed Methods Designs—Principles and Practices. Health Services Research. 2013;48(6 Pt 2):2134-56.
26. Onwuegbuzie A, Collins K. A Typology of Mixed Methods Sampling Designs in Social Science Research. The Qualitative Report. 2015.
27. Brooks J, McCluskey S, Turley E, King N. The Utility of Template Analysis in Qualitative Psychology Research. Qualitative Research in Psychology. 2015;12(2):202-22
28. Jonsson A. Facilitating productive use of feedback in higher education. Active Learning in Higher Education. 2013;14(1):63-76.
29. Harrison CJ, Könings KD, Schuwirth L, Wass V, van der Vleuten C. Barriers to the uptake and use of feedback in the context of summative assessment. Advances in Health Sciences Education. 2015;20(1):229-45.
30. ATLAS.ti Scientific Software Development GmbH. 22.0.11.0 ed.
31. Dey I. Grounding grounded theory: guidelines for qualitative inquiry. San Diego: Academic Press; 1999. 282 p.
32. Saunders B, Sim J, Kingstone T, Baker S, Waterfield

J, Bartlam B, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Quality & Quantity. 2018;52(4):1893-907.

33. Birt L, Scott S, Cavers D, Campbell C, Walter F. Member Checking: A Tool to Enhance Trustworthiness or Merely a Nod to Validation? Qualitative Health Research. 2016;26(13):1802-11.
34. Olmos-Vega FM, Stalmeijer RE, Varpio L, Kahlke R. A practical guide to reflexivity in qualitative research: AMEE Guide No. 149. Medical Teacher. 2022;0(0):1-11.
35. Orsmond P, Merry S. Feedback alignment: Effective and ineffective links between tutors' and students' understanding of coursework feedback. Assessment & Evaluation in Higher Education. 2011;36:125-36.
36. Handley K, Price M, Millar J. Beyond 'doing time': investigating the concept of student engagement with feedback. Oxford Review of Education. 2011;37(4):543-60.
37. Bing-You RG, Paterson J, Levine MA. Feedback falling on deaf ears: residents' receptivity to feedback tempered by sender credibility. Medical Teacher. 1997;19(1):40-4.
38. Winstone NE, Nash RA, Rowntree J, Menezes R. What Do Students Want Most from Written Feedback Information? Distinguishing Necessities from Luxuries Using a Budgeting Methodology. Assessment & Evaluation in Higher Education. 2016;41(8):1237-53.
39. Spooner M, Larkin J, Liew SC, Jaafar MH, McConkey S, Pawlikowska T. "Tell me what is 'better'!" How medical students experience feedback, through the lens of self-regulatory learning. BMC Medical Education. 2023;23(1):895.
40. Kulik JA, Kulik C-LC. Timing of Feedback and Verbal Learning. Review of Educational Research. 1988;58(1):79-97.
41. Bayerlein L. Students' feedback preferences: how do students react to timely and automatically generated assessment feedback? Assessment & Evaluation in Higher Education. 2014;39(8):916-31.
42. 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.
43. 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.
44. Carless D. Differing perceptions in the feedback process. Studies in Higher Education. 2006;31(2):219-33.
45. Hounsell D. Towards more sustainable feedback to students. Rethinking Assessment in Higher Education. 2007. pp. 101-13.
46. Heeneman S, Oudkerk Pool A, Schuwirth LWT, van der Vleuten CPM, Driessen EW. The impact of programmatic assessment on student learning: theory versus practice. Medical Education. 2015;49(5):487-98.
47. Prince KJAH, Boshuizen HPA, van der Vleuten CPM, Scherpbier AJJA. Students' opinions about their preparation for clinical practice. Medical Education. 2005;39(7):704-12.
48. Malau-Aduli BS, Roche P, Adu M, Jones K, Alele F, Drovandi A. Perceptions and processes influencing the transition of medical students from pre-clinical to clinical training. BMC Medical Education. 2020;20(1):279.
49. Burr SA, Brodier E, Wilkinson S. Delivery and use of individualised feedback in large class medical teaching. BMC Medical Education. 2013;13(1):63.
50. van Houten-Schat MA, Berkhout JJ, van Dijk N, Endedijk MD, Jaarsma ADC, Diemers AD. Self-regulated learning in the clinical context: a systematic review. Medical Education. 2018;52(10):1008-15.
51. Lucieer SM, Jonker L, Visscher C, Rikers RMJP, Themmen APN. Self-regulated learning and academic performance in medical education. Medical Teacher. 2016;38(6):585-93.
52. Carless D, Winstone N. Teacher feedback literacy and its interplay with student feedback literacy. Teaching in Higher Education. 2020:1-14.
53. Nouns ZM, Georg W. Progress testing in German speaking countries. Medical Teacher. 2010;32(6):467-70.

Appendix

Appendix 1 – Supplemental Table 1

Supplemental Table 1A. Feedback of the progress test with the results per category.

		Individual				Test moment group (*n*=57)							
Description categories	**Number of questions**	**Correct**	**Incorrect**	**?**	**Score**	**Correct**	**Std**	**Incorrect**	**Std**	**?**	**Std**	**Score**	**Std**
01 Respiratory system	13	69	31	0	56	68	13	28	12	4	7	57	18
02 Musculoskeletal system	17	59	41++	0--	38	58	11	31	9	11	10	44	14
03 Mental Health Care	16	69	31+	0	58-	75	14	20	12	5	7	68	18
04 Reproductive system	11	45-	55++	0	27--	58	15	29	13	13	13	48	18
05 Blood, lymph, heart and circulation	24	58	25	17+	48	60	13	29	11	11	9	48	17
06 Hormones and metabolism	13	46-	46++	8	-29	57	13	31	14	12	10	46	17
07 Skin and connective tissue	12	83	17	0	78	80	10	17	10	3	6	74	13
08 Personal, social and prevention aspects	17	29--	71++	0--	4--	52	14	38	14	11	10	35	19
09 Digestive system	17	71	29	0--	61	66	12	26	11	8	7	57	15
10 Kidneys and urinary tract	16	69	25	6	59	71	13	21	11	7	8	63	16
11 Nervous system and senses	17	47--	47++	-6	28--	62	13	26	12	12	11	53	16
12 Knowledge about skills	23	48	39	13	33	49	11	40	11	11	9	32	14
Total	196	57-	38++	5-	42-	62	8	29	6	9	6	51	9

-/--/++/+ low respectively high in comparison with the total group. Results are presented in percentages. Std = standard deviation. ? = question mark option use.

Supplemental Table 1B. Feedback of the progress test with the results per discipline.

		Individual				Test moment group (n = 57)							
Description disciplines	**Number of questions**	**Correct**	**Incorrect**	**?**	**Score**	**Correct**	**Std**	**Incorrect**	**Std**	**?**	**Std**	**Score**	**Std**
Anatomy	12	58	33	8	46	60	15	34	14	6	9	48	20
Biochemistry, molecular and cellular biology and genetics	18	50	44++	6--	34	46	14	31	12	24	14	34	17
Pharmacology	8	62	25	12	54	65	15	27	14	8	9	54	20
Physiology	11	73	27+	0-	62	73	17	18	12	9	12	65	21
Patho-, immuno- en microbiology	10	50	40	10	33-	57	15	34	15	10	10	44	19
Basic-, supportive subjects	**59**	**58**	**36+**	**7-**	**44**	**58**	**9**	**29**	**7**	**13**	**8**	**47**	**10**
Epidemiology/statistics	7	71+	29	0-	57+	55	23	32	15	12	21	41	26
Metamedica	5	20--	80++	0-	-23--	51	23	38	23	11	14	32	33
Psychiatry/psychology	12	67	33++	0-	54-	73	14	20	12	7	10	65	17
Social medicine	3	33	67+	0-	0	42	26	51	28	8	15	17	37
Behavioural scientific/other subjects	**27**	**56**	**44++**	**0-**	**35-**	**61**	**13**	**30**	**9**	**9**	**10**	**47**	**15**
Surgery	16	69	31	0-	56	67	13	27	12	6	8	56	17
Dermatology/ENT/ opthalmology	14	57	36	7	44	63	14	29	14	8	10	53	18
Geriatrics	8	62	38+	0	44	68	17	29	16	3	6	55	23
Obstetrics/Gynaecology	7	43--	57++	0-	21--	60	14	28	17	13	14	49	19
Family medicine	20	40--	55++	5	21--	61	12	34	12	4	5	49	16
Internal medicine	26	73	19	8+	67	73	11	22	9	5	5	64	14
Paediatrics	12	50-	42++	8	32-	60	15	28	13	12	12	48	19
Neurology	7	43	43+	14	19-	50	17	32	17	18	19	37	21
Clinical subjects	110	57-	37++	5	43--	65	8	28	7	7	6	54	10

-/--/++/+ low respectively high in comparison with the total group. Results are presented in percentages. Std = standard deviation. ? = question mark option use.

Appendix 2 – Interview guide

Part 1. Own feedback experiences
1.Do you prepare for the progress test?
How do you prepare? What determines whether you prepare for the progress test?
2. Do you consult the result of the progress test?
Which methods do you use to consult the test result? What determines whether you look at the test result?
3. Do you use the result of the progress test?
What do you do with this information? What determines whether you use the feedback?
4. Are you aware of the online feedback system (ProF)?
Why are you not using ProF? What do you think is the reason that you are not aware of ProF?
Part 2. Perception of progress test and feedback
1. What is your perception of the progress test? And which place does it have in your study program?
2. What is your perception of the way(s) the test result is presented to you?
Do you have any suggestions for improvements?

Appendix 3 – Reflexivity statement

The principal and second investigator (EvW, FvB) kept reflective diaries to create awareness of personal expectations, assumptions, and reactions to the participants and data. These diaries were used to guide critical dialogues during data analysis and clarify our interpretations of the data. During data collection, EvW experienced that she could easily relate to the participants, because of her own medical background and experience with the PT. This created an open atmosphere, in which the students felt comfortable to talk openly about their experiences and perceptions. Influenced by her scientific background in (bio) medicine EvW attempted to attain as much objectivity and produce rigorous qualitative research by using maximum variation sampling, member checking, and reflexivity throughout the data collection and analysis. The other researchers were an educational consultant and researcher in medical education (FvB) and a medical doctor with experience in clinical teaching and educational research (AL). Due to his background in cognitive psychology, FvB has been trained to conduct research in an empirical way, based on psychological theories. As such, he supported using theoretical concepts from feedback literature to formulate a priori themes. This theory-driven approach may have influenced the results. AL is a member of the national PT working group and a PT examiner, which might have influenced her perceptions on student behaviour.

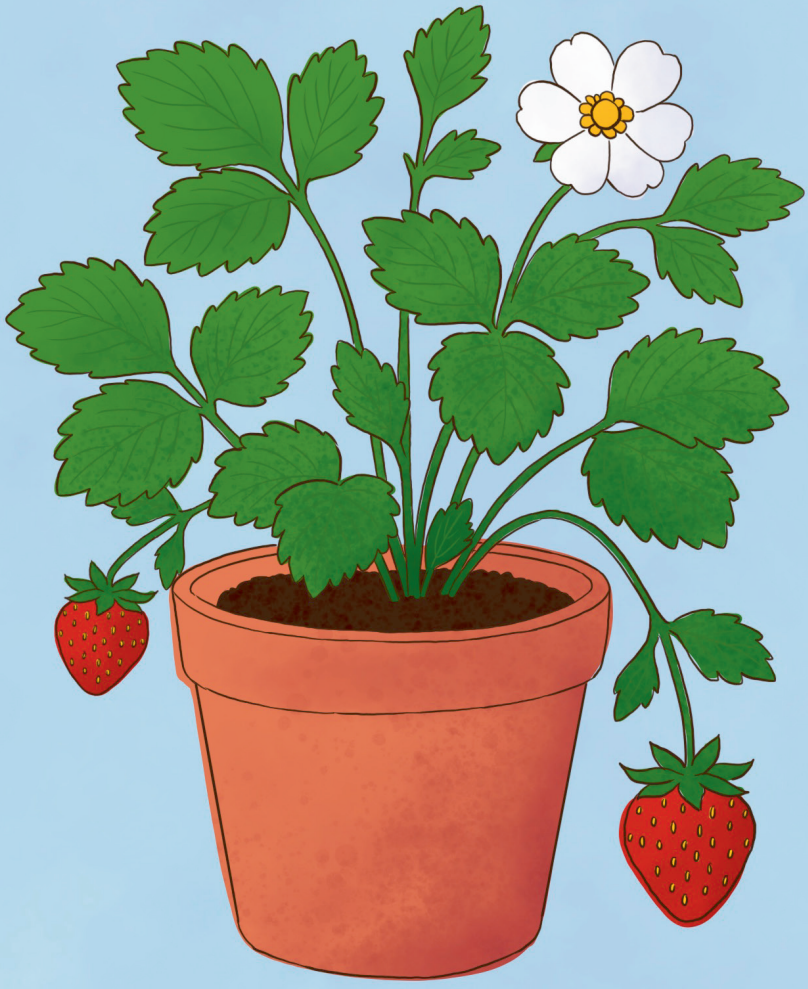
Appendix 4 – Supplemental Table 3

Supplemental Table 2. Descriptive characteristics of the participants

Number of logging sessions[a]	0		1		2 to 4		>5		
Grade	Fail	Pass/ Good	Fail	Pass/ Good	Fail	Pass/ Good	Fail	Pass/ Good	Total year (M/F)
Year 2	11[c]	3[c], 10[c]		21[c]		20[b],19[c]			6 (1/5)
Year 3	7[c]	18[c]	4[b]	9[c]		5[b], 8[b]	2[c]	15[b]	8 (4/4)
Year 5	6[c]		14[c]		16[c]	17[b]			4 (1/3)
Year 6	13[b]			1[c]		12[b]			3 (2/1)
Total Fail and Pass/Good (M/F)	4 (1/3)	3 (0/3)	2 (1/1)	3 (0/3)	1 (0/1)	6 (5/1)	1 (0/1)	1 (1/0)	

[a]ProF logging sessions from September 2020 to January 2021. M = male, F = female.

[b]Male. [c]Female.



Chapter 10

General discussion

General aim

Assessment is a key driver of student learning [1], making it a powerful tool to influence learning behaviour and foster ongoing development. Traditionally, assessment has been used primarily as a means of measuring knowledge and achievement (assessment *of* learning). However, its role has evolved toward an approach that fosters and facilitates learning (assessment *for* learning) [2, 3]. This shift highlights the growing need to design assessments that not only evaluate student performance, but also enhance engagement, support deeper learning, and promote self-regulation. While the potential of assessment to stimulate learning is widely acknowledged [1], further exploration is needed to optimize assessment design and maximize these benefits. Ensuring the effectiveness of such innovations requires alignment with established criteria — including ***acceptability, authenticity, catalytic effect, cueing effect, educational effect, equivalence, feasibility, reliability, testing effect,*** and ***validity*** [4-8]. In particular, principles that emphasize assessment *for* learning, such as the ***catalytic effect, testing effect,*** and ***educational effect*** are essential considerations in refining assessment strategies (see *Table 1* in the General introduction for the definition of the criteria).

This thesis addresses this gap by investigating how very short answer questions (VSAQs), computer adaptive progress testing, and feedback post-assessment can be leveraged to enhance student learning while improving assessment practices. Ultimately, this research aims to contribute to the development of assessment strategies that better prepare medical students for their future careers.

In this chapter, we provide the main findings, conclusions, and future research avenues for each of the three parts of this thesis separately: VSAQs, computer adaptive progress testing, and feedback. These conclusions are guided by the assessment development criteria introduced in the general introduction. After discussing each part individually, we then reflect on the practical implications of all our findings.

Part I: Very short answer question

Main findings

In the first study (**chapter 2**), we compared the reliability, discrimination, and acceptability of VSAQs and multiple-choice questions (MCQs) in a cohort of undergraduate medical students. Consistent with previous research [8], VSAQs demonstrated greater reliability and discriminative ability than MCQs in formative exams, with an acceptable average marking time of two minutes per question for the full student cohort. VSAQs were less susceptible to cueing effects than MCQs, but students reported greater uncertainty when answering VSAQs. Approximately half the students indicated that they would adjust their preparation strategy for this format. Additionally, most students perceived VSAQs to be more reflective of clinical practice.

In the following study (**chapter 3**), we investigated whether VSAQs or MCQs more effectively distinguished undergraduate medical students across different academic performance levels in summative medical examinations, as measured by grade point average (GPA). Across all three cohorts of first- and second-year students, student performance on VSAQs had a stronger positive association with GPA compared to MCQs. Moreover, VSAQs were overall better able to distinguish poor (i.e., lowest quintile of GPA) and excellent performing (i.e., highest quintile of GPA) students than MCQs.

The last study (**chapter 4**) explored the effectiveness of retrieval practice (i.e., testing effect) using MCQ and VSAQ practice tests on knowledge retention. We found no significant differences in knowledge retention between the two question formats and no interaction effect between question format on the practice and final test, suggesting that neither format is superior for fostering knowledge retention through retrieval practice. Our findings also demonstrated greater difficulty of VSAQs on both practice and final tests. Most students found the practice tests beneficial, regardless of the question format.

Conclusion

In conclusion, we investigated the effects of VSAQs on several key assessment criteria that contribute to high-quality assessment [4-8]. Our findings consistently support the psychometric advantages of VSAQs compared to MCQs, even when teachers have limited prior experience in VSAQ question design. VSAQs demonstrated high ***reliability*** and strong discriminative ability within formative assessments, aligning with previous research [8-10]. Unlike MCQs, VSAQs are less susceptible to ***cueing effects*** and guessing, reducing extraneous noise and enhancing their discriminative ability within individual examinations [8, 11-13]. Because they can more accurately differentiate students based on their understanding, they provide a more ***valid*** measure of the intended knowledge construct within an examination. Importantly, we demonstrated that this finding holds across multiple examinations, as VSAQs more effectively differentiated students based on GPA, further reinforcing the construct ***validity*** of assessments using VSAQs.

In terms of ***acceptability*** by teachers, we show that VSAQs can be marked efficiently within the digital assessment systems using a large cohort. This resonates with earlier findings regarding the marking time [8, 9]. However, the ***acceptability*** of this question format by students is just as, or even more, important for successful implementation. Even though students found the VSAQs more difficult and experienced more uncertainty while answering these questions, which could hamper their ***acceptability***, we suspect that due to the increased perceived ***authenticity*** of the question format over time this will be widely accepted by the students [14]. Although we did not explicitly study the ***educational effect*** (i.e., influence of question format on study behaviour), students reported they would prepare differently when assessed with VSAQs.

VSAQs are also highly suitable for practice tests, in which the open-ended nature provides valuable insights in students' misperceptions and knowledge gaps, which in turn can enhance the learning process (***catalytic effect***) [15, 16]. Although this question format requires more retrieval effort, which is beneficial for the ***testing effect*** [17], we did not observe improved knowledge retention compared to testing with MCQs. This lack of enhancement may be due to the lower initial retrieval success associated with VSAQs [18-21], which was evident in the lower practice test VSAQ scores. Since low initial retrieval success can weaken the ***testing effect***, it is important to balance retrieval effort and retrieval success to maximize the benefits of VSAQs in practice testing.

Future research avenues

A promising direction for future research is the integration of artificial intelligence (AI) to enhance both the grading and construction of open-ended questions, including VSAQs and essay questions. While VSAQs can efficiently assess clinical reasoning, essay questions may be better suited for evaluating more complex clinical reasoning and argumentation.

However, their adoption is particularly limited by time-intensive grading, even more so than VSAQs. Future research could explore how AI-driven grading models can address this challenge by improving both the accuracy and efficiency of reviewing open-ended responses [65]. Additionally, AI could support the generation of high-quality assessment questions that align with learning objectives.

Beyond undergraduate medical education, the application of VSAQs in postgraduate training and workplace-based assessments could be explored. Given their potential to better assess students' clinical reasoning than MCQs [42], VSAQs could serve as a more authentic assessment tool in these settings. Additionally, investigating whether VSAQs can predict future clinical performance would offer insights into their long-term validity as an assessment tool. Another promising direction for research is evaluating their role in student selection procedures, assessing whether VSAQs can better identify candidates likely to succeed in medical training.

Finally, while VSAQs align with key assessment criteria [4-8], their actual impact on student learning behaviour remains an open question. Evidence suggests that students employ more analytical reasoning when answering VSAQs [42], but further research is needed to determine their influence on deep learning. Longitudinal studies could explore how VSAQ-induced retrieval practice shapes learning behaviour, motivation, and learning strategies over time.

Part II: Computer adaptive progress testing

Main findings

In **chapter 6**, we investigated the correlation between student performance on a computer adaptive-progress test (CA-PT) and conventional progress test (PT) in nearly 1,500 medical students across different stages of study and medical schools in the Netherlands. We also assessed the feasibility of the CA-PT across medical schools. We observed a strong correlation between scores on the two PT formats (Pearson's r=0.83). The CA-PT was administered without technical issues and completed in a median time of 83 minutes (67–102 minutes). In the questionnaires, students reported perceiving the CA-PT as more challenging, but remained motivated to perform well.

In **chapter 7**, we explored the relationship between question mark option use in the conventional PT and performance on the CA-PT using retrospective data from nearly 6,000 medical students. In the conventional PT, the formula scoring method is applied, meaning incorrect answers result in penalties. However, students have the option to leave questions unanswered by selecting a question mark, which does not incur any penalty. Among students with similar conventional PT scores, those who frequently left questions unanswered (i.e., used the question mark option more often) generally performed better on the CA-PT, where a question mark option is lacking. However, this effect diminished as students progressed through their studies. To further examine the underlying structure of this relationship, we applied cluster analysis, which revealed a more nuanced pattern of variation between student subgroups within each study year. In year 4, student test-taking behaviour showed substantial variability, whereas in year 5 the pattern reversed — students who left more questions unanswered generally performed worse on the CA-PT. Additionally, we found a strong correlation between PT formats over time (Pearson's r=0.74).

Conclusion

To conclude, we demonstrate that the CA-PT is a ***reliable***, ***valid*** and efficient digital assessment format suitable for large-scale implementation across multiple medical schools. This personalized testing approach accommodates students at different stages of their studies without requiring formula scoring, which is necessary in the linear-fixed format of the conventional PT [22, 23]. Moreover, by removing the question mark option, the construct ***validity*** of the PT is enhanced, as our findings suggest that formula scoring may measure additional constructs — such as metacognitive skills and test-taking strategies — rather than solely knowledge. Consequently, the CA-PT allows for a more ***reliable*** assessment of students' knowledge levels.

Beyond its psychometric strengths, adaptive testing offers significant practical advantages that enhance the ***feasibility*** of the PT. Besides a shorter assessment duration, it provides greater flexibility, and improved scalability, since an established item bank removes the need to develop a new test for each administration and simultaneous administration across institutions is no longer necessary. Additionally, the ability to calibrate new questions during each test session streamlines item bank expansion, creating a self-sustaining system that reduces long-term resource demands. Nevertheless, implementing adaptive testing requires substantial initial investment and strong institutional collaboration to ensure its success [24].

A noteworthy implication of adaptive testing is that, by tailoring questions to students' knowledge levels, most items remain challenging regardless of ability. Our findings reflect this, as students reported encountering fewer questions they felt confident about on the CA-PT compared to the conventional PT. While this may reduce students' ability to gauge their performance during and after the test — potentially lowering self-efficacy and increasing anxiety [25] — it also reduces extraneous cognitive load by eliminating the need to decide which questions to answer first or how to navigate the test. As a result, students can focus more on answering questions rather than managing test-taking strategies, making the experience less cognitively demanding. Despite potential concerns about self-efficacy and uncertainty, students remained engaged and motivated, suggesting that, with proper preparation and clear communication of the test's purpose, the CA-PT may achieve high long-term student ***acceptability***.

Future research avenues

The implementation of CAT in medical education presents several opportunities for further research. While this dissertation has demonstrated the feasibility, reliability and validity of CAT in progress testing, future studies could expand on these findings to optimize its application and explore its broader impact.

One promising direction is the development of multidimensional CAT [66], which could enhance test precision by considering multiple parameters beyond student performance alone. Currently, unidimensional CAT adjusts question difficulty based solely on prior responses, but future research could investigate how incorporating factors such as item discrimination, and different subject domains could improve measurement accuracy. Another important area of research is the psychological impact of CAT on students. While our findings suggest that students remained motivated despite perceiving the CA-PT as more challenging, qualitative research could provide more in-depth insights into how adaptive testing influences self-efficacy, test anxiety, and perceived fairness over time.

Beyond progress testing, CAT could be explored in other assessment contexts, including formative practice tests and summative course assessments. Future research could evaluate its impact on learning behaviour, test-taking strategies, and long-term knowledge retention when used in different educational settings.

Part III: Feedback

Main findings

In **chapter 8** we applied the Expectancy Value Theory (EVT) [26] in a mixed-methods study to compare test preparation, feedback use, and test-taking motivation among medical students completing a purely formative PT versus a PT with a summative component (i.e., yielding of study credits). Students were more likely to consult feedback after the summative PT. However, test preparation, and active feedback use were relatively low and similar across both assessment conditions. Feedback engagement and test-taking motivation were influenced by the perceived value of the assessment. Performance-oriented students viewed the formative PT as unimportant due to absence of study credits, leading to low effort and limited feedback use. In contrast, learning-oriented students valued the formative PT for self-study and self-assessment, utilizing the feedback to gain insights into their learning and knowledge gaps.

In the qualitative study of **chapter 9**, we investigated the processes and factors affecting medical students' feedback use within the context of the Dutch PT, guided by Winstone *et al.*'s framework for effective feedback use [27]. Most students struggled to understand the feedback, were unaware of strategies and opportunities to use it effectively, felt disempowered or insecure when translating feedback into action, and lacked interest in the feedback. Several factors contributed to the perceived difficulties, such as the limited time, late timing of feedback, and unclear feedback presentation, and further hindered effective feedback use. However, feedback engagement increased during clinical rotations, where students sought feedback to better understand their performance levels and career prospects.

Conclusion

In conclusion, our findings demonstrate that the ***catalytic effect*** of the PT on student learning is currently limited, consistent with earlier studies [28-33]. Although the PT aims to promote reflection, identification of knowledge gaps, and ongoing learning through feedback [29], its perceived value is reduced when it does not yield study credits. This is especially true for performance-oriented students, who place less importance on assessments without direct study consequences, leading to reduced test-taking motivation and minimal feedback engagement. Notably, both performance- and learning-oriented students only actively engaged with feedback after failing the summative test. While grade focus tends to reduce feedback engagement once a satisfactory grade is achieved [27, 34, 35], the absence of urgency in the formative setting leads to even lower engagement. More broadly, this suggests that, for many students, the ***acceptability*** and ***catalytic effect*** of formative assessments are generally low when not directly linked to tangible rewards.

Furthermore, students' low engagement with PT feedback may stem from challenges with internal psychological processes essential for effective feedback use, such as awareness, cognizance, agency, and volition [27]. Our findings align with the established theoretical framework of Winstone *et al.* [27], and their recurrence in our study adds evidence to their importance and suggests that such difficulties may

be common across different educational contexts. We also identified specific factors — such as limited time, delayed feedback, and unclear presentation — that further hinder feedback engagement. While some of these factors are unique to our context, most resonate with prior research on feedback in other educational settings [36-39], underscoring their broader significance. Moreover, these context-specific factors present actionable targets for enhancing feedback engagement in similar educational settings.

Finally, our results reveal promising opportunities to enhance the ***catalytic effect*** of the PT, particularly among learning-oriented students and those in the clinical phase, where feedback engagement increased due to greater interest in performance and career prospects. Although engagement was mainly limited to feedback consultation, the clinical phase offers a key moment to strengthen feedback use, as students take a more serious approach to addressing knowledge gaps and applying knowledge in practice. Importantly, these findings suggest that fostering a sense of relevance and future applicability in the preclinical phase could help mitigate the low engagement in formative assessments. If students view feedback as a continuous developmental tool rather than something isolated to individual assessments, they may engage with it more meaningfully. This underscores the need to frame formative assessments in ways that highlight feedback as a lifelong learning skill rather than just an immediate tool for improvement.

Future research avenues

The transition from the conventional PT to the CA-PT presents new opportunities for research into student engagement with feedback in an adaptive assessment environment. Given that the CA-PT does not provide students with direct access to test questions post-examination but instead offers brief descriptions and external resources for further study, qualitative studies could explore how this change impacts feedback engagement, interpretation, and application.

Another interesting area for future research is the impact of feedback design and different assessment structures on student engagement. Our study identified barriers such as unclear feedback presentation, delayed delivery, and limited perceived relevance, all of which hinder effective feedback use. Future studies could explore whether real-time or more structured feedback mechanisms enhance feedback engagement. Additionally, examining whether these enhancements foster greater self-regulated learning, metacognitive development, and long-term retention would provide valuable insights into optimizing feedback's catalytic effect on student learning.

Implications for educational practice

Implement VSAQs in both formative and summative assessments

For decades, medical education has primarily relied on MCQs due to their high reliability, efficiency, and ease of grading [40, 41]. However, our findings in **chapter 2** highlight the psychometric advantages of VSAQs, which demonstrate higher ***reliability***, better item discrimination, and reduced ***cueing effects*** compared to MCQs. Given these benefits, integrating VSAQs into both formative and summative assessments can significantly enhance the quality and ***validity*** of medical assessments. Nevertheless, while incorporating VSAQs is highly beneficial, no single assessment method can fully capture all essential knowledge and competencies in medical education [40]. The most effective assessment strategies therefore combine various question formats, tailored to specific learning outcomes, Bloom's taxonomy levels, and the relevance of topics for students' future medical careers.

VSAQs should be used alongside other written question formats, and workplace-based assessments to ensure a comprehensive evaluation of students' knowledge, skills, and clinical reasoning abilities.

Use VSAQs in summative assessments to improve validity and differentiate student performance

Several studies, including our own studies (**chapter 2, 3**), highlight the superior discriminative ability of VSAQs within individual examinations [8-10]. VSAQs eliminate ***cueing*** and guessing, allowing scores to more accurately reflect students' true understanding and providing a more ***valid*** measure of knowledge within an examination. This results in a stronger ability to differentiate between high- and low-performing students. Beyond improving individual examination quality, VSAQs also provide a more accurate measure of student performance over time, distinguishing between different levels of academic performance level (i.e., GPA) across multiple examinations (**chapter 3**). This allows for the early identification of underperforming students who need additional support while simultaneously challenging high-achieving students. Additionally, using clinical vignettes for VSAQs enhances their ***authenticity***, aligning assessments more closely with real-world clinical reasoning and better preparing students for future practice [15, 42].

Implement VSAQs in formative assessments to familiarize students with the format and enhance learning

To enhance the ***acceptability*** of VSAQs, it is important to introduce them early in the curriculum. Integrating VSAQs into formative assessments throughout the curriculum allows students to become familiar with the question format, develop confidence, and gain valuable insights into their knowledge gaps and misconceptions. Regular exposure may also help reduce students' uncertainty about how to answer VSAQs compared to MCQs (**chapter 2**). Ensuring transparency in grading can further increase student acceptance — for example, clarifying that all answers will be reviewed and minor spelling errors will not be penalized may help reduce concerns about fairness.

Beyond familiarization, integrating VSAQs into formative assessments supports retrieval practice (i.e., the ***testing effect***), a well-established strategy for enhancing long-term retention [47]. Since our findings (**chapter 4**) did not demonstrate a clear advantage of VSAQs over MCQs for knowledge retention, the selection of the practice question format should be guided by the learning objectives. However, VSAQs reduce foresight bias, provide deeper insight into students' knowledge gaps, and promote conceptual understanding [15, 42, 48]. To optimize the initial low retrieval success of VSAQs and thereby its effectiveness, spaced retrieval practice and self-assessment with immediate self-feedback can help students recognize their knowledge gaps and refine their recall abilities [16, 49]. Additionally, supplementing retrieval practice with targeted restudy opportunities after receiving feedback may further enhance learning effectiveness [50]. A hybrid approach combining VSAQs with MCQs can effectively leverage the strengths of both formats [48, 51]. While VSAQs foster active recall and encourage deeper retrieval practice, MCQs offer the benefit of automated scoring, enabling students to receive immediate and corrective feedback. By integrating these formats, teachers can create a balanced approach to retrieval practice that optimizes both retrieval effort and success. Consequently, this combined approach supports improved formative assessment, facilitating greater long-term knowledge retention.

Support teachers in implementing VSAQs by providing training and addressing practical concerns

Successful implementation of VSAQs requires adequate preparation, not only for students but also for teachers. When introducing a new question format, it is essential to explain the rationale behind its selection and highlight its advantages in relation to the course's specific learning objectives. Providing targeted training for teachers, such as workshops, can equip teachers with the necessary skills to develop effective VSAQs while addressing potential practical concerns [43]. By clearly communicating the benefits of VSAQs — including the advantage that they do not require the creation of plausible alternative answer options, which can be challenging for MCQs [44-46] —teachers can be encouraged to integrate them into their assessments. Additionally, addressing concerns regarding the grading workload by demonstrating efficient review strategies can lower barriers to adoption. Alleviating practical concerns — in particular explaining how VSAQs can be reviewed in an acceptable amount of time — can lower the threshold for adoption and increase ***acceptability***. Seeking feedback from colleagues in different domains can provide diverse perspectives, further refining the effectiveness of VSAQs in medical education [43].

Implement CAT on a large-scale to assess students at different stages of their study

CAT presents an innovative approach to large-scale assessment, particularly in progress testing. As demonstrated in our study (**chapter 6**), CAT is a ***reliable***, efficient, and ***feasible*** test format that tailors assessments to students at different stages of their study. Unlike the conventional fixed-linear PT, CAT dynamically adjusts question difficulty based on student performance, eliminating the need for formula scoring while ensuring an accurate measurement of knowledge [52]. Beyond its use in assessment, the extensive data generated through CAT presents valuable opportunities for student progress monitoring, curriculum development, and educational research.

Prepare students to the transition to CAT by providing information and practice opportunities

To ensure successful implementation, students must be adequately prepared for this new adaptive test format. Our findings (**chapter 6**) indicate that students initially perceived CAT as more challenging and felt uncertain about their performance. Institutions should address these concerns by offering clear explanations, Q&A sessions, and practice tests that allow students to experience the format firsthand. Additionally, emphasizing the advantages of CAT, such as its ability to provide precise evaluations and tailored feedback, can help improve student ***acceptability*** and confidence.

Establish a collaborative approach to maintain a high-quality CAT item bank

While the transition to CAT requires a significant initial investment, its long-term benefits — such as improved flexibility and ***reliability*** — make it a valuable advancement in medical education [24]. However, a well-functioning CAT system depends on a large, high-quality item bank that continuously evolves to align with the curriculum. Given the substantial resources required for item development, collaboration among multiple medical schools is essential. By sharing expertise, test items, and validation efforts, institutions can ensure a steady supply of well-calibrated questions, improving the ***reliability*** and fairness of assessments while reducing individual institutional burdens.

Ensure that the chosen scoring method aligns with the assessment's goals

Our findings (**chapter 7**) indicate that formula scoring assesses not only knowledge but also metacognitive awareness and risk tendencies, which can impact the construct validity of test scores.

Therefore, if the primary goal is to measure knowledge in students with different knowledge levels, formula scoring may not be the appropriate scoring method. However, if the aim is to assess metacognitive awareness [53], formula scoring could be justified. Alternatively, self-assessment methods such as certainty-based marking (CBM) [54], may offer a more effective approach by incorporating students' confidence levels into the assessment, thereby enhancing both accuracy and self-reflection [55, 56].

Embed feedback as an integral part of the learning process

To maximize the ***catalytic effect***, feedback should be integrated into the medical curriculum and designed to actively engage students in their own learning process. Ensuring that students understand, value, and apply feedback requires a shared commitment from students, teachers, and program coordinators [57, 58]. This collaborative responsibility is essential in developing students' feedback literacy and fostering a culture where feedback is seen as a tool for continuous improvement rather than an evaluative measure.

Design structured and accessible feedback to enhance engagement

To enhance feedback engagement, both the format and delivery of feedback should be carefully designed by teachers and course coordinators. Providing clear instructions, specific feedback messages, and timely access to feedback can help overcome common student barriers, such as uncertainty about how to interpret and apply the feedback (**chapter 9**) [27, 36-39]. Well-structured feedback allows students to engage more effectively and take meaningful action based on their knowledge gaps.

Integrate progress test feedback into course activities to enhance relevance

Ensuring meaningful engagement can be particularly challenging for curriculum-independent assessments, such as the PT, as students struggle to recognize its relevance (**chapter 9**). To strengthen the impact of PT feedback, it should be embedded into course learning activities [39, 57, 59, 60]. One effective strategy might be to discuss PT feedback in small-group discussions with mentors, where students can reflect on their feedback, ask questions, and develop concrete learning strategies. Additionally, aligning course practice questions with PT content by teachers can help students recognize its connection to their coursework, making PT feedback a more natural and integrated part of their learning process. A well-integrated approach allows students to engage with feedback, identify and address challenging areas, and reinforce learning through immediate application.

Support students in understanding and using feedback

While self-regulation is increasingly emphasized in medical education [61, 62], our findings (**chapter 9**) indicate that many students do not use the opportunities available to reflect on their feedback. This suggests that explicit guidance remains essential for effective feedback use. To foster meaningful engagement, feedback literacy should be developed early in the curriculum and through interactive dialogues with teachers [58, 63], with a strong emphasis on its long-term value beyond assessments. In our study, students perceived formative assessment as less important compared to summative assessments, which reduced its perceived importance and feedback engagement (**chapter 8**). To address this, teachers should clearly communicate the purpose of formative assessments and highlight how feedback supports ongoing learning and professional growth. For example, in programmatic assessment [2], feedback is embedded into continuous learning rather than treated as an isolated event, helping students recognize its value and integrate it into their learning process.

General conclusion

Our research highlights that assessment in medical education is not merely a measurement tool, but a fundamental driver of student learning. By optimizing assessment strategies through innovative approaches such as VSAQs, CAT, and structured feedback, we can substantially enhance student learning, leading to improved knowledge retention, skill development, and preparedness for professional practice. Our findings offer valuable insights for refining written assessments, aligning them closely with established criteria for high quality assessments. Specifically, implementing VSAQs into medical curricula improves the validity and authenticity of assessments, while CAT provides more individualized and reliable assessments. Embedding feedback as an integral part of the learning process can foster a culture that values formative assessment, motivating students to engage actively with and benefit from feedback. Ultimately, integrating these complementary innovations offers a robust approach to assessment, ensuring medical education supports student growth and lifelong learning.

References

1. Newble DI, Entwistle NJ. Learning styles and approaches: implications for medical education. Medical Education. 1986;20(3):162-75.
2. Schuwirth LWT, Van der Vleuten CPM. Programmatic assessment: From assessment of learning to assessment for learning. Medical Teacher. 2011;33(6):478-85.
3. Scott IM. Beyond 'driving': The relationship between assessment, performance and learning. Medical Education. 2020;54(1):54-9.
4. 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.
5. Gulikers JTM, Bastiaens TJ, Kirschner PA. A five-dimensional framework for authentic assessment. Educational Technology Research and Development. 2004;52(3):67-86.
6. Roediger HL, Butler AC. The critical role of retrieval practice in long-term retention. Trends in Cognitive Sciences. 2011;15(1):20-7.
7. 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.
8. Sam AH, Field SM, Collares CF, van der Vleuten CPM, Wass VJ, Melville C, et al. Very-short-answer questions: reliability, discrimination and acceptability. Medical Education. 2018;52(4):447-55.
9. 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.
10. 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.
11. Schuwirth LWT, Vleuten CPM, Donkers HHLM. A closer look at cueing effects in multiple-choice questions. Medical Education. 1996;30(1):44-9.
12. 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.
13. 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.
14. Puthiaparampil T, Rahman MM. Very short answer questions: a viable alternative to multiple choice questions. BMC Medical Education. 2020;20(1):141.
15. 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.
16. 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).
17. 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.
18. Smith MA, Karpicke JD. Retrieval practice with short-answer, multiple-choice, and hybrid tests. Memory. 2014;22(7):784-802.
19. Moreira BFT, Pinto TSS, Starling DSV, Jaeger A. Retrieval Practice in Classroom Settings: A Review of Applied Research. Frontiers in Education. 2019;4.
20. 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.
21. McDermott KB, Agarwal PK, D'Antonio L, Roediger HL, McDaniel MA. Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. Journal of Experimental Psychology: Applied. 2014;20(1).
22. 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.
23. Lord FM. Formula scoring and number-right scoring. Journal of Educational Measurement. 1975;12(1):7-11.
24. 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.
25. Martin AJ, Lazendic G. Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. Journal of Educational Psychology. 2018;110(1):27-45.
26. 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.
27. 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.Van Berkel HJM, Nuy HJP, Geerligs T. The influence of progress tests and block tests on study behaviour. Instructional Science. 1994;22(4):317-33.

28. 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.
29. Aarts R, Steidel k, Manuel BAF, Driessen EW. Progress testing in resource-poor countries: A case from Mozambique. Medical Teacher. 2010;32(6):461-3.
30. 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.
31. 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.
32. 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):Doc41.
33. Carless D. Differing perceptions in the feedback process. Studies in Higher Education. 2006;31(2):219-33.
34. Hounsell D. Towards more sustainable feedback to students. Rethinking Assessment in Higher Education. 2007:101-13.
35. Kulik JA, Kulik C-LC. Timing of Feedback and Verbal Learning. Review of Educational Research. 1988;58(1):79-97.
36. Bayerlein L. Students' feedback preferences: how do students react to timely and automatically generated assessment feedback? Assessment & Evaluation in Higher Education. 2014;39(8):916-31.
37. Cordovani L, Tran C, Wong A, Jack SM, Monteiro S. Undergraduate Learners' Receptiveness to Feedback in Medical Schools: A Scoping Review. Medical Science Educator. 2023;33(5):1253-69.
38. Lefroy J, Watling C, Teunissen PW, Brand P. Guidelines: the do's, don'ts and don't knows of feedback for clinical education. Perspectives on Medical Education. 2015;4(6):284-99.
39. Al-Rukban MO. Guidelines for the construction of multiple choice questions tests. Journal of Family & Community Medicine. 2006;13(3):125-33.
40. Schuwirth L, van der Vleuten C. Written Assessment. ABC of Learning and Teaching in Medicine: Wiley-Blackwell. 2017; 65-9.
41. 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.
42. van Wijk EV, Janse RJ, Langers AMJ. Response to: 'Twelve tips for introducing very short answer questions (VSAQs) into your medical curriculum'. Medical Teacher. 2023;45(5):553-4.
43. Little JL, Bjork EL. Optimizing multiple-choice tests as tools for learning. Memory & Cognition. 2015;43(1):14-26.
44. Little JL, Frickey EA, Fung AK. The role of retrieval in answering multiple-choice questions. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2019;45(8).
45. Gierl MJ, Bulut O, Guo Q, Zhang X. Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review. Review of Educational Research. 2017;87(6).
46. 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).
47. van den Broek GSE, van Gog T, Jansen E, Pleijsant M, Kester L. Multimedia Effects During Retrieval Practice: Images That Reveal the Answer Reduce Vocabulary Learning. Journal of Educational Psychology. 2021;113(8):1587-608.
48. Vaughn KE, Rawson KA, Pyc MA. Repeated retrieval practice and item difficulty: Does criterion learning eliminate item difficulty effects? Psychonomic Bulletin & Review. 2013;20(6).
49. Storm BC, Friedman MC, Murayama K, Bjork RA. On the transfer of prior tests or study events to subsequent study. Journal of experimental psychology Learning, memory, and cognition. 2014;40(1).
50. Park J. Learning in a New Computerized Testing System. Journal of Educational Psychology. 2005;97(3).
51. Chang H-H. Psychometrics behind Computerized Adaptive Testing. Psychometrika. 2015;80(1):1-20.
52. Krathwohl DR. A Revision of Bloom's Taxonomy: An Overview. Theory Into Practice. 2002;41(4).
53. Gardner-Medwin AR. Confidence assessment in the teaching of basic science. Research in Learning Technology. 1995;3(1).
54. Cash B, Mitchner NA, Ravyn D. Confidence-Based Learning CME: Overcoming Barriers in Irritable Bowel Syndrome With Constipation. Journal of Continuing Education in the Health Professions. 2011;31(3).
55. Luetsch K, Burrows J. Certainty rating in pre-and post-tests of study modules in an online clinical pharmacy course - A pilot study to evaluate teaching and learning. BMC Medical Education. 2016;16(1).
56. Molloy E, Boud D, Henderson M. Developing a learning-centred framework for feedback literacy. Assessment & Evaluation in Higher Education. 2020;45(4):527-40.

57. Carless D, Winstone N. Teacher feedback literacy and its interplay with student feedback literacy. Teaching in Higher Education. 2020:1-14.
58. 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.
59. Watling CJ, Ginsburg S. Assessment, feedback and the alchemy of learning. Medical Education. 2019;53(1):76-85.
60. van Houten-Schat MA, Berkhout JJ, van Dijk N, Endedijk MD, Jaarsma ADC, Diemers AD. Self-regulated learning in the clinical context: a systematic review. Medical Education. 2018;52(10):1008-15.
61. Lucieer SM, Jonker L, Visscher C, Rikers RMJP, Themmen APN. Self-regulated learning and academic performance in medical education. Medical Teacher. 2016;38(6):585-93.
62. Ajjawi R, Regehr G. When I say … feedback. Medical Education. 2019;53(7):652-4.
63. Bloom BS. Taxonomy of Educational Objectives: The Classification of Educational Goals: Longmans, Green; 1956. 240 p.
64. Grévisse C. LLM-based automatic short answer grading in undergraduate medical education. BMC Medical Education. 2024;24(1).
65. Wang C, Weiss DJ, Su S, Suen KY, Basford J, Cheville A. Multidimensional Computerized Adaptive Testing: A Potential Path Toward the Efficient and Precise Assessment of Applied Cognition, Daily Activity, and Mobility for Hospitalized Patients. Archives of physical medicine and rehabilitation. 2022;103(5 Suppl).

Appendices

Nederlandse samenvatting
List of scientific contributions
Curriculum Vitae
Dankwoord

Nederlandse samenvatting

Toetsing speelt een cruciale rol in het medisch onderwijs; het dient niet alleen om kennis te meten, maar het kan ook leergedrag van studenten sturen. Schriftelijke tentamens worden veelvuldig toegepast binnen medische curricula om op een betrouwbare manier kennisniveaus te beoordelen. Deze toetsen bestaan voornamelijk uit meerkeuzevragen en worden vooral gebruikt om studentprestaties te meten aan de hand van cijfers. Er is echter een verschuiving gaande in de perceptie van toetsing: de focus verschuift van toetsing als beoordelingsinstrument (summatieve toetsen) naar toetsing als leermiddel (formatieve toetsen). De huidige technologische ontwikkelingen bieden nieuwe mogelijkheden om bestaande toetsmethoden te verbeteren en nieuwe innovatieve methoden te ontwikkelen die het leren stimuleren. De implementatie van deze innovatieve toetsmethoden in het medisch onderwijs en de complexe relatie tussen toetsing en leren zijn echter nog onvoldoende onderzocht.

Dit proefschrift richt zich op de mogelijkheden en uitdagingen van verschillende innovaties binnen toetsing met als overkoepelend doel het optimaliseren van toetsing en leren van medisch studenten, zodat zij zich tot competente dokters kunnen ontwikkelen. Allereerst worden de mogelijkheden van een nieuw vraagformat, de *very short answer question* (VSAQ), onderzocht. Vervolgens bestuderen we de implementatie en voordelen van de computer adaptieve voortgangstoets, een toets die vier keer per jaar plaatsvindt en studenten inzicht geeft in hun kennisontwikkeling gedurende de opleiding. Tot slot onderzoeken we de verschillen in feedbackgebruik van studenten tussen formatieve en summatieve toetsen en de factoren die dit feedbackgedrag beïnvloeden. We hebben de effectiviteit van de verschillende innovaties onderzocht aan de hand van vastgestelde criteria voor kwalitatief goede toetsen: aanvaardbaarheid, authenticiteit, cueing effect, educatief effect, haalbaarheid, katalyserend effect, testeffect en validiteit.

Deel I: Very short answer question

Meerkeuzevragen worden in het medisch onderwijs veelvuldig gebruikt vanwege hun betrouwbaarheid en de efficiëntie waarmee ze nagekeken kunnen worden. Echter, ze hebben ook beperkingen, omdat studenten bij het leren vooral op herkenning vertrouwen in plaats van op begrip van de leerstof. Daarnaast kunnen studenten gebruik maken van aanwijzingen (*cues*) in de antwoordopties om bij het juiste antwoord te komen of kunnen ze gokken. VSAQ's kunnen deze nadelen omzeilen. Dit vraagformat vraagt om zeer korte, open antwoorden van slechts één tot vier woorden, wat studenten stimuleert om actief kennis op te halen en zelf een antwoord te generen op basis van hun kennis. Het voordeel van VSAQ's ten opzichte van uitgebreide open vragen is dat ze makkelijker toepasbaar en sneller objectief na te kijken zijn, wat met name van belang is in grote groepen studenten. Ondanks deze voordelen van VSAQ's is er nog weinig onderzoek gedaan naar de betrouwbaarheid en toepasbaarheid van VSAQ's in het medisch onderwijs, vooral in contexten waarin studenten en docenten nog weinig ervaring hebben met dit vraagformat.

In **hoofdstuk 2** beschrijven we een studie waarin we eerdere bevindingen over de betrouwbaarheid (de mate waarin de vraag steeds opnieuw hetzelfde, nauwkeurige resultaat oplevert), het onderscheidend vermogen (de mate waarin de vraag een onderscheid maakt tussen studenten met veel en weinig kennis over de leerstof) en de acceptatie van VSAQ's hebben onderzocht bij geneeskundestudenten en docenten die niet eerder met dit vraagformat werkten.

In deze studie werden studenten verdeeld in twee groepen voor een formatieve toets: de ene groep kreeg eerst VSAQ's en daarna meerkeuzevragen, de andere groep andersom. De resultaten tonen aan dat VSAQ's een hogere betrouwbaarheid en beter onderscheidend vermogen hebben dan meerkeuzevragen. Bovendien was de tijd die nodig was om de VSAQ's na te kijken acceptabel, namelijk gemiddeld twee minuten per vraag voor alle 375 studenten van een geheel cohort. Daarnaast kwam er ook minder '*cueing*' (i.e., het beantwoorden van de vraag door gebruik te maken van cues in de vraag of antwoordopties) voor in VSAQ's dan in meerkeuzevragen. Uit de vragenlijsten bleek ook dat studenten bij VSAQ's minder zeker waren van hun antwoorden. Daarnaast bereidde ongeveer de helft van de studenten zich anders voor op dit format en vond dat VSAQ's een betere afspiegeling van de klinische praktijk gaven. Hiermee bevestigen we dat VSAQ's ook buiten een ervaren onderzoeksgroep goed toepasbaar, betrouwbaar en authentiek zijn.

In **hoofdstuk 3** onderzochten we het onderscheidend vermogen van beide vraagformats verder, waarbij we ons hebben gericht op het onderscheidend vermogen over meerdere toetsen heen in plaats van binnen de toets. Uit eerder onderzoek blijkt dat een vraagformat met een sterk onderscheidend vermogen binnen een toets niet vanzelfsprekend ook een sterk onderscheidend vermogen over meerdere toetsen heen. We bestudeerden welk vraagformat, VSAQ of meerkeuzevraag, beter in staat is om studenten te onderscheiden op basis van hun academische prestatieniveau, gemeten als cijfergemiddelde. In deze retrospectieve studie werden de scores geanalyseerd van twee summative toetsten uit het eerste en tweede jaar van de medische bachelor, die bestonden uit zowel VSAQ's als meerkeuzevragen. De analyses werden uitgevoerd op drie studentcohorten, terwijl binnen elk cohort twee studentpopulaties werden onderscheiden: de eerste populatie bestond uit eerstejaarsstudenten die het eerstejaarstentamen hadden gemaakt, en de tweede populatie bestond uit tweedejaarsstudenten die zowel het eerste- als het tweedejaarstentamen in opeenvolgende jaren hadden afgelegd. De VSAQ's hadden een sterkere positieve associatie met het cijfergemiddelde dan meerkeuzevragen in alle cohorten. Daarnaast waren ze over het algemeen ook beter in staat om slecht en excellent presterende studenten te onderscheiden op basis van hun cijfergemiddelde. Hiermee tonen we aan dat VSAQ's beter in staat zijn om studenten met verschillende academische prestatie niveaus te identificeren. Het gebruik van VSAQ's in toetsen kan daarmee het vroegtijdig identificeren van zowel slecht als excellent presterende studenten verbeteren, waardoor er gepaste begeleiding of extra uitdaging kan worden aangeboden.

In **hoofdstuk 4** richten we ons op een andere waardevolle eigenschap van toetsen, namelijk hun gunstige effect op het vasthouden van kennis op de lange termijn (i.e., kennisretentie). Het ophalen van kennis uit het geheugen, bekend als '*retrieval practice*' of het '*testeffect*', is een bewezen effectieve strategie om leren te verbeteren en kennisretentie te vergroten. Deze strategie wordt vaak toegepast via oefentoetsen, maar het is nog onduidelijk welk vraagformat hierbij het meest effectief is. In een gerandomiseerde studie onderzochten we het effect van oefentoetsen met VSAQ's en meerkeuzevragen op de kennisretentie van studenten. Studenten in een extracurriculaire cursus over leefstijl beantwoordden zowel VSAQ's als meerkeuzevragen, zonder dat ze na afloop feedback kregen. Drie weken na de laatste oefentoets werd hun kennis getest met identieke vragen in beide formats in een eindtoets. Hoewel de VSAQ's moeilijker bleken en studenten hier lagere scores op haalden, wat doorgaans een positief effect heeft op kennisretentie, werd er geen verschil in retentie tussen de vraagformats gevonden. Dit betekent dat we geen bewijs hebben gevonden dat één van de vraagformats effectiever is in het bevorderen van kennisbehoud door middel van *retrieval practice*.

Een mogelijke verklaring voor het ontbreken van een verschil in effectiviteit is dat studenten bij de oefentoetsen lagere scores behaalden op de VSAQ's. Dit kan hebben geleid tot minder succesvol ophalen van kennis, waardoor de retentie op de eindtoets mogelijk ook lager uitviel. Om de effectiviteit van VSAQ's in *retrieval practice* te verbeteren en de kennisretentie te vergroten, kan het helpen om de slagingskans (het succes) bij het oefenen te verhogen. Dit kan bijvoorbeeld door directe feedback te geven en herhaalde oefenmomenten aan te bieden.

We sluiten deel I af met een opinie stuk in **hoofdstuk 5** over de rol van toetsing en vraagformats in het medisch onderwijs. We stellen hierbij de rol van de meerkeuzevraag als standaard vraagformat ter discussie. VSAQ's bieden namelijk een veelbelovend alternatief met meerdere voordelen: ze zijn betrouwbaarder, stimuleren actief leren en bereiden studenten beter voor op de klinische praktijk, waar patiënten zich ook niet met antwoordopties presenteren. Hoewel VSAQ's meer nakijktijd kosten dan meerkeuzevragen, blijkt dit relatief snel en efficiënt te kunnen met de huidige digitale toetsomgevingen. Daarnaast hoeven er voor dit vraagformat geen plausibele foutieve antwoordopties bedacht te worden zoals bij meerkeuzevragen, wat tijd bespaart in de vraagontwikkeling. Om de implementatie van VSAQ's in het medisch onderwijs soepel te laten verlopen, bevelen we onder andere workshops voor docenten en oefentoetsen voor studenten aan. Deze benadering helpt docenten ervaring op te doen met het ontwikkelen en nakijken van VSAQ's, terwijl studenten wennen aan het vraagformat en tegelijkertijd hun kennisniveau kunnen testen.

Deel II: Computer adaptieve voortgangstoets

De voortgangstoets (VGT) is een toets die vier keer per jaar door alle geneeskundestudenten wordt gemaakt om te volgen hoe hun kennis zich tijdens hun opleiding ontwikkelt. De toets bevat vragen uit alle vakgebieden van de opleiding en wordt steeds opnieuw aangeboden, waardoor studenten en docenten inzicht krijgen in wat goed beheerst wordt en wat nog extra aandacht vereist. Oorspronkelijk maakten alle studenten precies dezelfde toets, ongeacht hun studiejaar. Dit zorgde vooral bij jongerejaars studenten voor problemen, omdat zij regelmatig vragen kregen die te moeilijk waren voor hun kennisniveau, waardoor de toets minder betrouwbaar werd. *Computerized Adaptive Testing* (CAT) biedt hier mogelijk een oplossing, waarbij de moeilijkheid van de vragen wordt afgestemd op het individuele niveau van de student tijdens het maken van de toets. In **hoofdstuk 6** beschrijven we een cross-over studie waarin we de conventionele VGT vergeleken met een CAT-variant. Geneeskundestudenten van drie medische faculteiten maakten zowel de conventionele VGT als de CAT, waarna we de correlatie tussen de resultaten en de uitvoerbaarheid van de CAT onderzochten. De scores op de CAT bleken sterk te correleren met de scores op de conventionele VGT en de CAT werd zonder problemen afgenomen. De gemiddelde afnametijd werd met de CAT verkort tot 83 minuten, wat leidt tot een efficiëntere toets zonder daarbij aan betrouwbaarheid in te leveren. Deze positieve bevindingen hebben het invoeren van een landelijke CAT in het medisch curriculum ondersteund, waarbij een goede voorbereiding en voorlichting van studenten belangrijk is, aangezien studenten de CAT als uitdagender ervaren.

In de conventionele VGT werden punten afgetrokken voor incorrecte antwoorden. Studenten konden echter ook een vraagtekenoptie kiezen als ze het antwoord niet wisten, zonder dat dit tot puntenaftrek leidde. Deze scoringsmethode, bekend als '*formula scoring*', is bedoeld om gokken te ontmoedigen. Uit onderzoek blijkt echter dat het gebruik van de vraagtekenoptie sterk verschilt tussen studenten

en mogelijk samenhangt met risicogedrag en andere persoonlijke factoren. Met de introductie van de CAT ontstond een unieke situatie: studenten maakten zowel een VGT mét (conventioneel) als zonder (CAT) vraagtekenoptie. Hierdoor konden we op grote schaal onderzoeken hoe deze vraagtekenoptie de prestaties van studenten beïnvloedt. In **hoofdstuk 7** is de relatie onderzocht tussen het gebruik van de vraagtekenoptie in de conventionele VGT en de prestaties op de CAT. De resultaten laten zien dat er grote verschillen zijn in het gebruik van de vraagtekenoptie en het effect hiervan op de CAT-scores, zowel binnen als tussen studiejaren. Zo zagen we bijvoorbeeld dat in de eerste jaren van de studie het gebruik van meer vraagtekens, dus meer vragen onbeantwoord laten, samenhing met hogere scores op de CAT, terwijl dit effect in latere studiejaren minder werd en in het laatste jaar zelfs omkeerde waarbij meer vraagtekengebruik samenhing met lagere scores op de CAT. In jaar vier observeerden we de meeste variatie in de effecten van vraagtekengebruik op CAT scores tussen verschillende groepen studenten. Dit suggereert dat *formula scoring* niet alleen kennis meet, maar ook andere constructen, zoals invulgedrag en metacognitieve vaardigheden. Daarom is het belangrijk om de scoringmethode zorgvuldig te kiezen en af te stemmen op de toetsdoelen, zodat de resultaten eerlijk en betrouwbaar blijven. Als het doel van de toets puur het meten van het kennisniveau is, lijkt formula scoring hier niet geschikt voor.

Deel III: Feedback

Formatieve toetsen hebben als doel studenten inzicht te geven in hun sterke en zwakke punten door middel van uitgebreide feedback, zodat zij hun leerproces gericht kunnen verbeteren. Ondanks de toenemende implementatie van formatieve toetsen in het medisch onderwijs, blijkt het in de praktijk echter nog moeilijk om studenten optimaal gebruik te laten maken van de feedback. In **hoofdstuk 8** onderzochten we hoe studenten zich voorbereiden op formatieve en summatieve VGT's, hoe zij omgaan met de feedback na de verschillende toetsen en hoe hun motivatie verschilt wanneer een toets wel of niet meetelt voor studiepunten. In deze studie hebben we vragenlijstdata, inloggegevens van het online feedbacksysteem voor de VGT en interviewgegevens geanalyseerd om het feedbackgebruik en de motivatie voor de toetsen tussen de formatieve groep en de summatieve groep te vergelijken. Uit onze studie bleek dat studenten die de summatieve VGT maakten vaker de feedback raadpleegden, maar de voorbereiding en het actief gebruik van de feedback voor het eigen leerproces waren over het algemeen laag en vergelijkbaar in beide groepen. Onze interviewdata toonden aan dat er een onderscheid gemaakt kon worden tussen prestatiegerichte en leergerichte studenten. Prestatiegerichte studenten waardeerden de formatieve VGT minder vanwege het gebrek aan studiepunten en besteedden daarom minder aandacht aan deze toets en de bijbehorende feedback. Leergerichte studenten daarentegen zagen de formatieve VGT wel als een kans om hun kennis te evalueren en verbeteren en waren gemotiveerd om deze toets serieus te maken en de feedback te bekijken. Deze bevindingen suggereren dat studenten over het algemeen minder gemotiveerd zijn om zich in te zetten voor een toets en de bijbehorende feedback wanneer er geen directe (studie)consequenties aan verbonden zijn. Een leeromgeving die het belang van formatieve toetsing duidelijk benadrukt en ondersteunt is daarom essentieel om studenten te motiveren actief gebruik te maken van feedback in hun leerproces.

De studie in **hoofdstuk 9** onderzocht via interviews welke processen en factoren het feedbackgebruik van studenten in de context van de VGT beïnvloeden. Studenten gaven aan dat ze de feedback vaak moeilijk konden interpreteren, dat deze te laat werd gegeven en onvoldoende specifiek was.

Ook misten zij begeleiding in het toepassen van de feedback, waardoor veel studenten alleen actie ondernamen na een onvoldoende. Er werd meer gebruik gemaakt van de feedback in de coschappen, wanneer studenten een serieuzere houding ontwikkelden ten opzichte van hun studie en carrière perspectieven. Om het feedbackgebruik te stimuleren, bevelen we aan om feedback helder en tijdig beschikbaar te maken en om begeleiding aan te bieden, zodat de VGT niet alleen als meetinstrument, maar ook als leermiddel effectiever kan worden ingezet. Bovendien kan een betere integratie van de VGT en de feedback in het curriculum de betrokkenheid van studenten vergroten en hen beter voorbereiden op het toekomstige beroepsleven.

Op basis van de bevindingen in dit proefschrift geven we enkele praktische aanbevelingen voor het onderwijs:

Gebruik VSAQ's in zowel summatieve als formatieve toetsen
VSAQ's bieden een betrouwbaardere manier om kennis te toetsen dan meerkeuzevragen. Ze voorkomen gokken en geven een nauwkeuriger beeld van het werkelijke kennisniveau van studenten. Door het gebruik van VSAQ's kan de kwaliteit van toetsen worden verbeterd. Bovendien kunnen ze door hun goede onderscheidende vermogen docenten helpen bij het vroegtijdig identificeren van studenten die extra ondersteuning nodig hebben. Om studenten te laten wennen aan dit vraagtype, kan het nuttig zijn om VSAQ's al vroeg in het curriculum in formatieve (oefen) toetsen in te zetten. Zo raken studenten vertrouwd met het vraagtype, kunnen ze hun kennishiaten ontdekken en zich beter voorbereiden op summatieve toetsen. Om de kennisretentie te vergroten (het zogenaamde testeffect), is het zinvol om de oefentoetsen te combineren met feedback, zodat studenten direct inzicht krijgen in hun kennisniveau en zwakke punten. Daarnaast is het raadzaam om docenten goed voor te bereiden en te ondersteunen, bijvoorbeeld via interactieve workshops waarin zij leren hoe ze goede VSAQ's opstellen.

Implementeer CAT op grote schaal om studenten in verschillende stadia van hun studie te toetsen
CAT is een innovatieve toetsmethode waarbij de moeilijkheidsgraad van de vragen zich tijdens de toets automatisch aanpast aan het niveau van de student. Uit ons onderzoek blijkt dat CAT betrouwbaar, efficiënt en praktisch haalbaar is voor voortgangstoetsen, omdat studenten op verschillende momenten in hun studie nauwkeurig kunnen worden getoetst. Om CAT succesvol in te voeren, raden we aan om studenten duidelijk voor te bereiden door middel van informatiesessies, oefenmogelijkheden en heldere communicatie over de voordelen van CAT. Hoewel de invoering van CAT een investering vraagt, zijn de voordelen op lange termijn — zoals verbeterde betrouwbaarheid en flexibiliteit — zeer waardevol. Daarnaast is het belangrijk dat verschillende onderwijsinstellingen samenwerken om een uitgebreide, actuele en kwalitatief sterke vragenbank te ontwikkelen en te onderhouden.

Maak feedback een integraal onderdeel van het leerproces
Feedback kan een sterk positief effect hebben op leren en zou daarom een vanzelfsprekend onderdeel moeten zijn van het leerproces van studenten. Hiervoor is het belangrijk dat studenten begeleiding krijgen bij het effectief inzetten van feedback voor hun eigen ontwikkeling. Goede integratie van feedback vraagt duidelijke instructies, goede toegankelijkheid en regelmatige aandacht voor feedback binnen het curriculum. Vooral bij toetsen die minder direct aansluiten bij het actuele onderwijs, zoals voortgangstoetsen, kan het zinvol zijn om feedback actief te bespreken in kleine groepen met een docent.

Op die manier kunnen studenten hun resultaten beter begrijpen, vragen stellen en leerstrategieën ontwikkelen. Het aanbieden van aanvullende oefenvragen die aansluiten bij de inhoud van voortgangstoetsen helpt studenten bovendien om de verbinding te zien tussen feedback en hun onderwijs. Zo wordt feedback een effectief hulpmiddel voor studenten om hun leerproces te verbeteren.

Conclusie

Toetsing is meer dan alleen een meetinstrument; het is een sterke drijfveer voor leren. Door innovatieve methoden zoals VSAQ's, CAT en feedback, wordt toetsing niet alleen betrouwbaarder, maar ook een waardevol leermiddel voor studenten. VSAQ's verbeteren de kwaliteit van toetsen en stimuleren authentiek leren, terwijl CAT zorgt voor een efficiëntere, nauwkeurigere en meer gepersonaliseerde toetsing. Structureel geïntegreerde feedback kan studenten bovendien motiveren om actief met hun eigen leerproces aan de slag te gaan en zich te blijven ontwikkelen. Door deze verschillende methoden te combineren kunnen we toetsing betrouwbaarder en betekenisvoller maken voor studenten, zodat zij optimaal worden voorbereid op hun toekomstige carrière.

List of Scientific Contributions

Scientific publications

van Wijk EV, Janse RJ, Langers AMJ. Response to: 'Twelve tips for introducing very short answer questions (VSAQs) into your medical curriculum'. *Medical Teacher*. 2023;45(5):553-334. doi:10.1080/0142159X.2022.2158070.

van Wijk EV, Janse RJ, Ruijter BN, Rohling JHT, van der Kraan J, Crobach S, *et al.* Use of very short answer questions compared to multiple choice questions in undergraduate medical students: An external validation study. *PLoS ONE*. 2023;18(7):e0288558. doi:10.1371/journal.pone.0288558.

van Wijk EV, van Blankenstein FM, Donkers J, Janse RJ, Bustraan J, Adelmeijer LGM, Dubois EA, Dekker FW, Langers AMJ. Does 'summative' count? The influence of the awarding of study credits on feedback use and test-taking motivation in medical progress testing. *Advances in Health Science Education*. 2024;29:1665-1688. doi:10.1007/s10459-024-10324-4.

van Wijk EV, van Blankenstein FM, Janse RJ, Dubois EA, Langers AMJ. Understanding students' feedback use in medical progress testing: A qualitative interview study. *Medical Education*. 2024;58(8):980-988. doi:10.1111/medu.15378.

van Wijk EV, Janse RJ, Langers AMJ. Leren van Toetsen: Hoe je met goede toetsvragen het leren kunt stimuleren. *Nederlands Tijdschrift voor Geneeskunde*. 2024;168:D8012.

van Wijk EV, Donkers J, De Laat, PCJ, Meiboom AA, Jacobs B, Ravesloot JH, Tio RA, Van Der Vleuten CPM, Langers AMJ, Bremers AJA. *Perspectives on Medical Education*, 2024;13(1):406-416. doi:10.5334/pme.1345.

van Wijk EV, de Jonge M, van Blankenstein FM, Janse RJ, Langers AMJ. The battle of question formats: a comparative study of retrieval practice using very short answer questions and multiple choice questions. *BMC Medical Education*. 2024;24:1547. doi:10.1186/s12909-024-06538-0.

Manuscripts under review

van Wijk EV, Donkers J, De Laat PCJ, Meiboom AA, Jacobs B, Ravesloot JH, Tio RA, Oud FMM, Kooman JP, Bremers AJA, Langers AMJ. The effect of the question mark option in progress testing: A large scale longitudinal study. *Resubmitted after revisions to Perspectives on Medical Education*.

van Wijk EV, van Blankenstein FM, Ruijter BN, Rohling JHT, van der Kraan J, Dekker FW, Langers AMJ. Identifying academic success and underperformance: The discriminative power of very short answer questions and multiple-choice questions. *Submitted to Medical Teacher*.

van Wijk EV, van Blankenstein FM, Langers AMJ. Bridging assessment and clinical practice: The added value of very short answer questions in medical education. *A version of this manuscript has been accepted in The Clinical Teacher.*

Conference presentations and workshops

Format	Title	Year	Conference
Workshop	An introduction to the community of medical education	2022	LUMC Onderwijsconferentie
Workshop	VSAQs in digital examinations	2022	Leiden Education Festival
Presentation	Prior knowledge activation	2023	LEARN LUMC
Workshop	Hoe schrijf je een goede VSAQ?	2023	NVMO
Presentation	Feedback gebruik na een formatief en summatief afgenomen voortgangstoets geneeskunde: is er een verschil?	2023	NVMO
Presentation	Feedback use in a formative and summative medical progress test: A mixed-methods study	2023	AMEE
Presentation	Computer adaptive testing in progress testing: Feasibility and comparison with non-adaptive testing	2023	AMEE
Workshop	Hoe je met goede toetsvragen het leren kunt stimuleren	2024	LEARN LUMC
Presentation	Verbetering van het begrip van feedbackgebruik in de context van de medische voortgangstoets	2024	NVMO
Presentation	De computer adaptieve voortgangstoets vs. de conventionele voortgangstoets: prestaties, ervaringen en uitvoerbaarheid	2024	NVMO
Round table	Tussen droom en werkelijkheid voor de poort van de studie geneeskunde	2024	NVMO
Presentation	Understanding students' feedback use in medical progress testing	2024	AMEE
Presentation	De "strijd" der vraagtypes: een vergelijkende studie naar retrieval practice met very short answer questions en meerkeuzevragen	2025	NVMO

Curriculum Vitae

Elise Vivianne van Wijk werd geboren in Delft op 11 februari 1994. Zij kwam in een warm gezin terecht met twee oudere broers. Ze groeide op in Rijswijk. In 2012 behaalde zij haar gymnasium diploma cum laude aan het Gymnasium Haganum te Den Haag. Hierna begon zij aan de studie Biomedische Wetenschappen aan de Universiteit van Leiden. In haar tweede jaar heeft zij deelgenomen aan het uitwisselingsprogramma met het Karolinska Instituut in Stockholm, Zweden. Na het doorlopen van de bachelor, besloot zij haar studie te combineren met Geneeskunde via het zij-instroomprogramma. Na een introductieprogramma met essentiële onderdelen van de vakken van jaar 2 van de bachelor Geneeskunde, volgde zij het reguliere onderwijsprogramma van jaar 3. Alvorens haar master te starten, volgde zij een klinische stage bij de chirurgie, gynaecologie en de spoedeisende hulp in Kumasi, Ghana. Haar master bestond uit een vierjarig gecombineerd programma van de opleidingen Biomedische Wetenschappen en Geneeskunde (het “Health” traject), waarin zij naast de verplichte coschappen, twee onderzoeksstages heeft afgerond bij de psychiatrie en neurologie in het Leids Universitair Medisch Centrum (LUMC). Daarnaast heeft zij een keuzecoschap ouderen psychiatrie bij Parnassia en een semi-arts stage interne geneeskunde bij het Reinier de Graaf ziekenhuis gedaan. Haar opleiding werd kortdurend onderbroken door de COVID-19 pandemie, waarin zij tijdelijk als basisarts in een verpleeghuis heeft gewerkt. In 2020 behaalde zij haar diploma voor zowel Geneeskunde als Biomedische Wetenschappen, waarna zij aan het werk ging als arts niet in opleiding (ANIOS) op de interne geneeskunde in het Reinier de Graaf ziekenhuis. In september 2021 startte ze haar promotietraject in het medisch onderwijs bij het Onderwijs Expertise Centrum van het Leids Universitair Medisch Centrum onder leiding van promotors Prof. dr. A.M.J. Langers en Prof. dr. F.W. Dekker en copromotor dr. F.M. van Blankenstein. Gedurende dit promotietraject was zij ook werkzaam als toetsredacteur bij het Nederlands Tijdschrift voor Geneeskunde, waarbij zij toetsen maakt voor de nascholing van artsen. Tevens was zij lid van de voortgangstoets beoordelingscommissie (VBC), het jonge onderzoekers netwerk (JON) van de Nederlandse Vereniging van Medisch Onderwijs (NVMO) en onderdeel van de blokcommissie van de halve minor Medical Education. Ook was zij gedurende twee jaar secretaris van de LUMC Association for PhD Candidates (LAP), waarbij zij opkwam voor de belangen van de PhD-studenten van het LUMC en verschillende (netwerk) activiteiten organiseerde. In november 2023 onderbrak zij haar promotietraject voor twee maanden om een mooie reis te maken door Nieuw Zeeland met de camper. In april 2025 rondde zij haar promotietraject af. Inmiddels is zij werkzaam als ANIOS Bedrijfsgeneeskunde bij MKBasics.

Dankwoord

In april 2021 ontving ik een e-mail van Alexandra Langers over een PhD-vacature in het medisch onderwijs. Door mijn eerdere betrokkenheid bij de voortgangstoets dacht ze dat dit iets voor mij was. Ze sloot af met: "*Ik kan me voorstellen dat je misschien verrast bent door dit bericht*". En dat was ik zeker. Ons eerste gesprek voelde direct goed en een week later had ik een nieuwe baan bij het Onderwijs Expertise Centrum.

Tijdens mijn promotietraject werd ik omringd door lieve mensen die op hun eigen manier hebben bijgedragen. Ik wil in het bijzonder mijn begeleidingsteam bedanken. Alexandra, vanaf dag één voelde ik mij op mijn gemak en kreeg ik de ruimte om te groeien. Ondanks je drukke agenda vond je altijd tijd voor mijn manuscripten en onze overleggen. Jouw heldere blik, betrokkenheid en positieve energie waren van onschatbare waarde. Friedo, bedankt voor de leerzame (statistiek)overleggen, je aanmoediging en je bereidheid om altijd mee te denken. Floris, jouw kennis, engelengeduld en doortastendheid hebben me enorm geholpen. Je begeleidde me door het ongemak van kwalitatief onderzoek en we vierden onze publicaties met een biertje.

Mijn VSAQ-onderzoeken waren niet mogelijk zonder de steun van de blokcoördinatoren Jos Rohling, Bastian Ruijter en Jolein van der Kraan. Bij jullie is het "OVKA" onderzoek begonnen en ik hoop dat jullie je enthousiasme blijven verspreiden! Mario en Roemer, mijn OVKA-buddies, bedankt voor de fijne samenwerking, discussies en jullie enthousiasme.

Mijn lieve collega's van het OEC-onderzoeksteam wil ik ook bedanken. Lieve Marchien, jouw warme welkom, positieve energie en wijze adviezen maakten mijn start ondanks de COVID-lockdown zoveel fijner. Je werd opgevolgd door nog zo'n topper: Caroline. Je nuchtere blik en daadkracht maakten jou een echte aanwinst voor het team. Bedankt voor je steun in de laatste fase van mijn promotietraject. Lieve Jacqueline, AJ, Caroline en Esmée, het scholierenproject was een leuk en leerzaam avontuur: samen het land doorreizen, honderden vragenlijsten verwerken en lachen om alle gekke tekeningen.

Lieve PhD buddies: Jolande, Alice, Rozita en Dani, wat ben ik blij dat jullie aan mijn zijde stonden en wat hebben we een leuke tijd samen gehad! Jullie steun en waardering betekenden veel voor me en ik ben supertrots op jullie. V7-38 verlaat ik, maar jullie nog lang niet! Renée en Charlotte, bedankt voor jullie gezelligheid en hulp in mijn beginperiode. Lieve Lottes, wat was het fijn dat jullie ook in het LUMC werkten en we samen het PhD leed konden delen (het liefst in de zon).

Mijn collega's van de voortgangstoets, bedankt voor de samenwerking. Jeroen, je statistische kennis was erg inspirerend en ik heb veel van je geleerd. Maartje, zonder jou had ik dit promotietraject wellicht nooit gedaan. Jouw energie, creativiteit en doorzettingsvermogen bewonder ik enorm. Eline, bedankt voor je kritische blik, gezelligheid en oprechte interesse.

Lieve Laura, wat ben ik blij dat we elkaar ontmoet hebben tijdens de JON-dag. Samen startten we een journal club én een mooie vriendschap. Dansen in Glasgow, dobberen door Bazel en ontbijten in Egmond – wat wordt onze volgende bestemming?

Mijn paranimfen, lieve Robin en Britt, wat ben ik blij dat jullie naast me staan. Britt, als doorgewinterde PhD'er had je altijd een luisterend oor of goed advies. Dankzij jou ontdekte ik de elektrische fiets en maakten we veel gezellige ritjes. Onze LUMC-tijd is voorbij, maar onze vriendschap blijft nog heel lang bestaan! Robin, mijn grote broer, je bent altijd al een voorbeeld geweest, maar de afgelopen jaren zijn we nog closer geworden. Heel bijzonder om deze mijlpaal met je te delen (en hopelijk heb ik hiermee mijn eerdere keuze goedgemaakt).

Lieve Eefie, onze reis door Nieuw-Zeeland voelde als een ware sabbatical. Twee maanden samen in een camper, genietend van de rust en natuur – precies wat ik nodig had om de laatste loodjes door te komen!

Lieve Nicole, mijn allerbeste vriendin. Onze koffiemomentjes, lunches en wandelingen maakten mijn promotietraject zoveel dragelijker. Ik waardeer enorm dat je er altijd voor me bent en we alles met elkaar kunnen delen! Ik kan me geen betere girlfriend wensen.

Lieve Eva, ontzettend bedankt voor je hulp met de lay-out. Jouw oog voor detail en creatieve ideeën hebben er een prachtig proefschrift van gemaakt.

Lieve Dorus, van data analyse tot InDesign, niks was je te gek. Werken op de Paleisstraat was altijd een feestje met liters koffie en pindakaas binnen handbereik. Bedankt voor de fijne gesprekken, je geduld, vertrouwen en eindeloze steun.

Lieve Clemens, mijn andere grote broer. We hebben al veel meegemaakt samen, wat onze band hecht heeft gemaakt. Ik ben trots op je doorzettingsvermogen en positiviteit. Lieve papa en mama, zonder jullie steun was ik nooit zo ver gekomen. Jullie gaven me de basis om te groeien en stimuleren me om door te zetten. Bedankt voor jullie vertrouwen, betrokkenheid en liefde.

En tot slot, allerliefste Isar. Jouw onvoorwaardelijke steun en liefde betekenen alles. Je weet me altijd op te vrolijken en geeft me de moed om mijn grenzen te verleggen. Bij jou kan ik helemaal mezelf zijn. Ik kijk uit naar alles wat de toekomst ons nog gaat brengen!