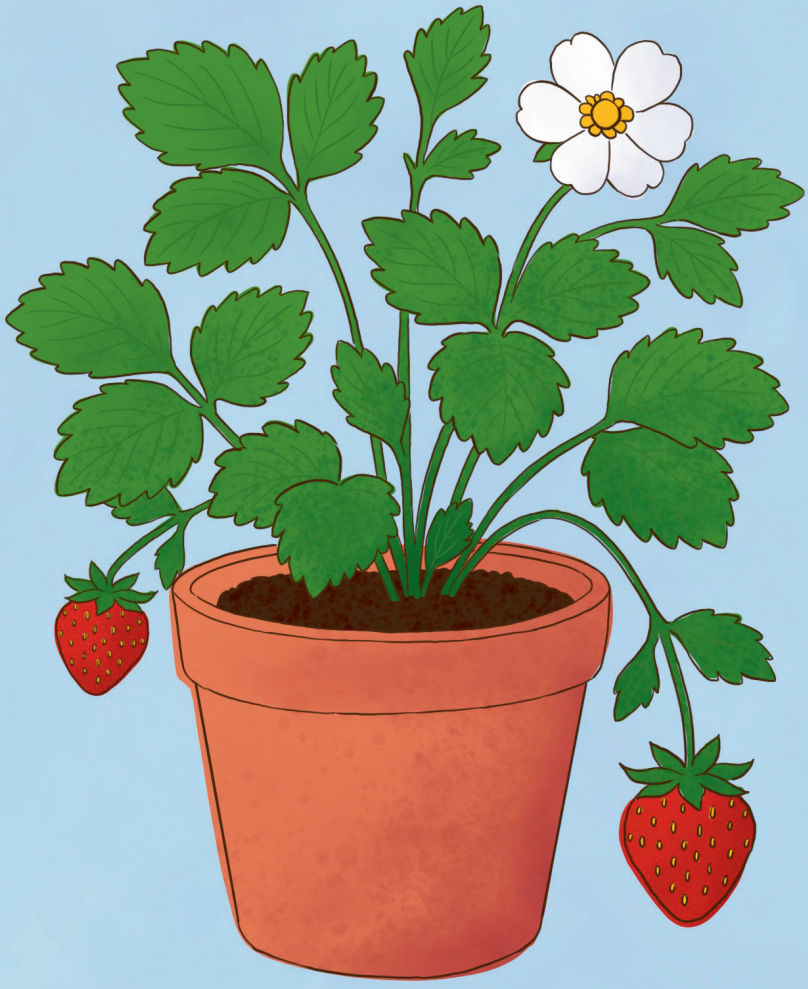
Assessment for growth: fostering student learning through assessment innovations in medical education

Wijk, E.V. van



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).