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

Wijk, E.V. van

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*).

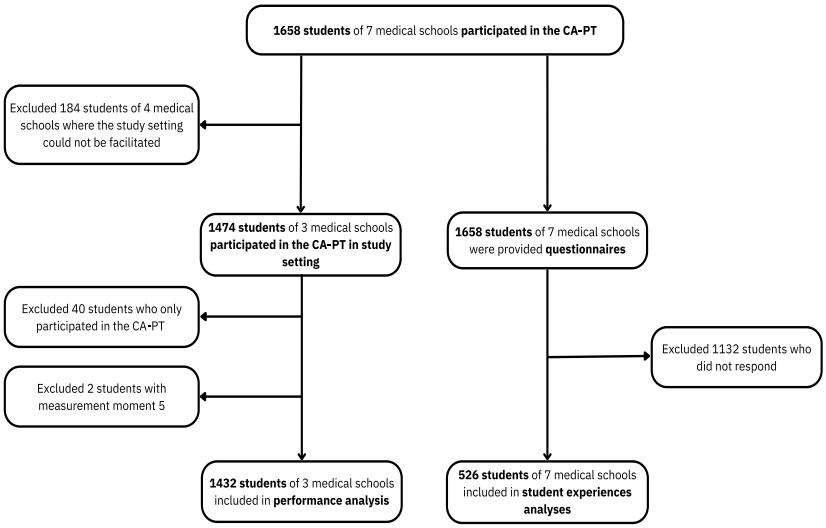

Figure 1. Flowchart of participants in questionnaire and test performance analyses

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*).

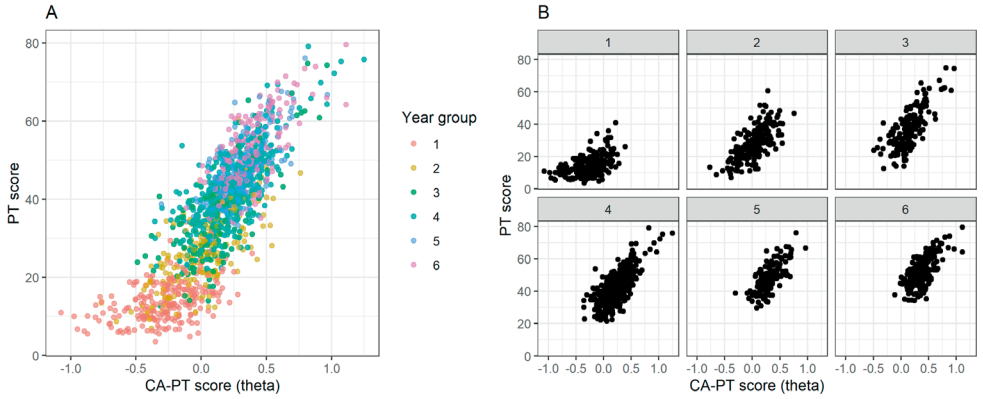

Figure 2. The relationship between the z-scores on the conventional PT (y-axis) and the theta on the CA-PT (x-axis) for ***(A)*** year 1 to 6 and ***(B)*** each year separately.

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).

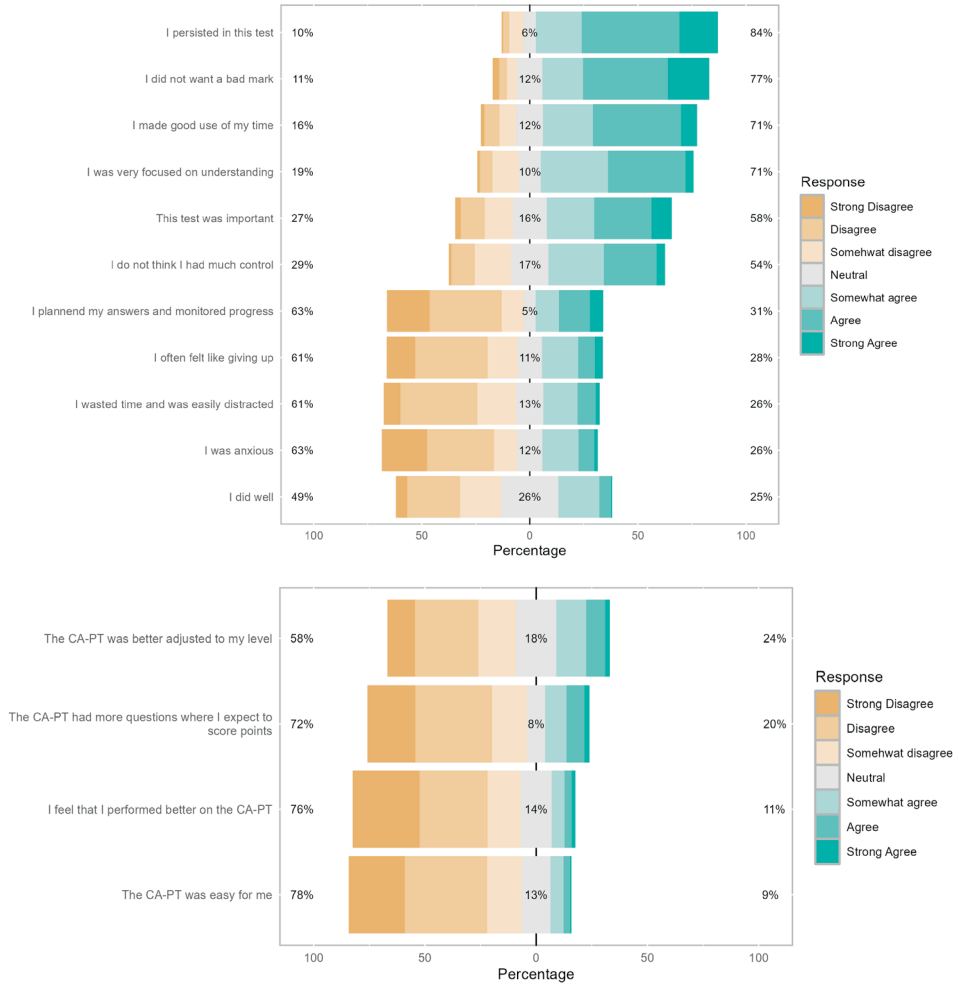

Figure 3. Distribution of answers to the questionnaire items ***(A)*** 1 to 11 on motivation and engagement in the computer adaptive progress test; ***(B)*** 12 to 15 comparing the conventional progress test with the computer adaptive progress test. I persisted in this test = "*I persisted in this test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