

2023-10-25

Medical students' perceptions of a novel international adaptive progress test

Kisielewska, J

<https://pearl.plymouth.ac.uk/handle/10026.1/21487>

10.1007/s10639-023-12269-4

Education and Information Technologies

Springer

All content in PEARL is protected by copyright law. Author manuscripts are made available in accordance with publisher policies. Please cite only the published version using the details provided on the item record or document. In the absence of an open licence (e.g. Creative Commons), permissions for further reuse of content should be sought from the publisher or author.



Medical students' perceptions of a novel international adaptive progress test

Jolanta Kisielewska^{1,2} · Paul Millin^{1,2} · Neil Rice^{2,3} ·
Jose Miguel Pego^{2,4,5} · Steven Burr^{1,2} · Michal Nowakowski^{2,6} ·
Thomas Gale^{1,2}

Received: 26 May 2023 / Accepted: 6 October 2023
© The Author(s) 2023

Abstract

Between 2018–2021, eight European medical schools took part in a study to develop a medical knowledge Online Adaptive International Progress Test. Here we discuss participants' self-perception to evaluate the acceptability of adaptive vs non-adaptive testing. Study participants, students from across Europe at all stages of undergraduate medical education with varying levels of prior experience with progress testing, sat remotely invigilated tests using the online QuizOne® platform. Participants completed online feedback questionnaires on their experiences and perceptions of adaptive and non-adaptive tests. Overall satisfaction with the organisation and delivery of remote online tests was high regardless of previous experience with progress testing, differences in stages, programmes, and to some degree language. In statements probing the appropriateness of the level and the length of testing, differences were observed between adaptive and non-adaptive tests. There was a high level of agreement that the adaptive test was a good measure of personal knowledge and increased participants' motivation for study. Students' self-perception of the assessment is an important factor in evaluation of acceptability of the exam and its further development. In our study, the adaptive test algorithm adjusted the level of difficulty for the individual student in real-time, leading to positive perceptions of the length of the test and promoting students' engagement. The assessment increases student motivation for learning and in turn, has the potential to improve their performance.

Keywords Computerised adaptive testing · Progress testing · Self-perception · Remote examination · International medical students

Extended author information available on the last page of the article

Published online: 25 October 2023

Springer

1 Introduction

Effective and efficient delivery of healthcare education depends on reliable assessment that promotes deep learning and prepares prospective doctors to practice. The use of longitudinal progress testing can be very effective in steering students' learning and monitoring their progress against final programme outcomes (Ali et al., 2016, 2018).

In most medical schools in Europe, the knowledge of healthcare students is assessed in a compartmentalised curriculum, which may promote superficial rather than life-long learning habits (Devine et al., 2015). To foster self-directed learning and to assess the development of student knowledge, progress testing was introduced over 40 years ago in healthcare programs (Van der Vleuten et al., 1996). It relies on a longitudinal assessment strategy in which students' knowledge is tested periodically using assessments composed of content from the whole curriculum (Wrigley et al., 2012). The success of progress testing relies mainly on the type of feedback students receive, impacting on self-perception of this assessment. Therefore, many educators and assessors aim for a better alignment between tested knowledge and students' individual needs for learning. Computerised adaptive testing (CAT) allows adaptation of test difficulty to an individual learner based on an algorithm that selects test questions dependent on the previous responses of a test-taker (Collares & Cecilio-Fernandes 2019). It is considered that CAT represents a significant evolution in assessment methods (Wainer et al., 2000).

One of the purposes of any low-stakes assessment for example in a programmatic assessment framework is to ensure that candidates can obtain sufficient feedback that they can use to improve learning. If the assessment provides an appropriate level of feedback, learners may change their behaviour accordingly to engage with the learning goals. Therefore, the concept of self-perception is incredibly important to link the assessment type and learners' experience with motivation for learning (Mohebi & Bailey, 2020). Thus, the purpose of developing adaptive progress testing for medicine was to ensure that students, irrespective of their previous experience, are motivated to increase their own learning and achieve the required level of knowledge to be safe practitioners. Determining whether an assessment method is fit for purpose, both in terms of its implementation and sustainability, relies heavily on psychometric analysis and students' evaluation (Boud & Soler, 2015). It has been shown that adaptive quizzes can increase students' motivation and engagement in learning based on immediate identification of knowledge gaps (Ross et al., 2018). Adaptive progress testing may not necessarily improve students' performance in the short term (Griff & Matter, 2013) but over the longer-term permits personalised learning opportunities between tests, and more reliably determines the level of knowledge attained by each student irrespective of their country or programme (Rice et al., 2022).

In 2018–2021, eight medical schools from five countries across Europe contributed to an ERASMUS + funded project to develop an Online Adaptive International Progress Test (OAIPT), (Rice et al., 2022). Using demographic and

self-perception feedback collected during this project, the current aim is to evaluate student participants' experience and satisfaction with remotely delivered non-adaptive and adaptive progress tests. Because the perception of the exam drives its successful implementation, we focus here on evaluation of the level of student acceptability for adaptive vs non-adaptive testing and whether there were any differences in student perception between countries, curricula and stages.

2 Overview of the development of Online Adaptive International Progress Test project

Eight medical schools from five countries across Europe (Finland, Netherlands, Poland, Portugal and UK), all members of the European Board of Medical Assessors (EBMA), contributed to an ERASMUS+ funded project to develop the OAIPT. The project included the development of a large bank of 1000+ English language questions covering international medical curricula and the delivery of three tests, two non-adaptive and one adaptive. Non-adaptive progress tests were used to calibrate the item bank, and performance data were analysed to determine the validity and difficulty parameters for each item. A third test was delivered using a Computerized Adaptive Testing (CAT) algorithm. High test reliability for all participants irrespective of the country and stage ensured the validity across the study population (Rice et al., 2022).

To address the acceptability of the technology and test modality, demographic and self-perception surveys were delivered at the end of each test.

3 Method

3.1 Student recruitment and administration of tests

Undergraduate medical students from all stages (years 1–6) were recruited for the project. Emails, flyers, and advertisements were distributed explaining the rationale of the project. The process of student recruitment was repeated before the first pilot and subsequent test administration. Students who responded were provided with a Participant Information Sheet, a consent form, and details of each stage of testing. The consent form was signed at the point of delivery of each test.

The first non-adaptive test was delivered online using Test Life®, a bespoke test delivery platform hosted by Maastricht University, in face-to-face settings with invigilators physically present in the room. Due to the COVID-19 pandemic, the second non-adaptive test and the adaptive test (Test 1 and Test 2 respectively for our analyses) were also delivered online, but remotely using the QuizOne® platform with proctors present virtually.

For this study, we focused on the survey data from the tests delivered remotely using the QuizOne® platform in order to maintain a uniform student experience (Tests 1 and 2).

Each of the tests (Tests 1 and 2) contained 125 items. Psychometric evaluation of the tests showed very high reliability for both the non-adaptive and adaptive tests. Average conditional individual reliability being 0.892 and relatively stable across the midrange of the latent person ability scale for the non-adaptive, and above 0.9 for the adaptive test and stable across the full range of participant abilities with adaptive tests converging to stable ability parameters with low measurement error in around a third the length of non-adaptive tests (Rice et al., 2022).

3.2 Participant survey development

The project partners jointly developed a questionnaire to capture information about student demographics, language proficiency, and self-perception of the exam oriented towards an active student role in the quality of the assessment. The statements were designed based on the model by Brown and Hirschfeld, so the students could participate fully in developing an assessment that is enjoyable, engaging, and motivating for further learning (Brown & Hirschfeld, 2008). Participating students were invited to complete an evaluation at the end of each test.

Socio-demographic variables were included in order to profile the sample of participating students in line with the German Social Science Infrastructure Services (GESIS) Guidelines (Hoffmeyer-Zlotnik, 2016). English language proficiency questions were based on the Common European Framework of Reference for Languages (CEFR, Levels and Europe, 2018). The self-perception statements were focused on the evaluation of various aspects of the exam including instructions and help; delivery of the exam; software; quality of questions; and structure of the exam. In addition to the general survey delivered for Test 1 and 2, the adaptive test (Test 2), had five additional statements addressing perceptions on how well the adaptive test was aligned with the student's knowledge and how the test motivated students further learning. Test-related statements were developed using a Likert scale from 1 to 5, where 1 was the lowest (fully disagree), and 5 was the highest evaluation (fully agree). A Likert scale from 1–5 is a commonly used scale when studying data on self-reported satisfaction of participants (Ho, 2017; Sullivan & Artino, 2013) and is appropriate for use post-assessment delivery to facilitate fast responses to survey statements.

3.3 Questionnaire analysis

Two analyses were carried out, the first analysis examined the difference in levels of agreement between Tests 1 and 2 taken at two different times by different participants. The second analysis examined the level of agreement with a different set of statements specifically related to perceptions of the adaptive test (Test 2 only).

3.3.1 Analysis one

The principal area of interest was participants' perceptions of differences between non-adaptive (Test 1) and adaptive (Test 2) tests. In particular, analysis was

undertaken on the responses to statements 7, 8, 9, and 10 of the survey (statements content included in Table 3), as these were the declarations that probe the underlying differences between the adaptive and non-adaptive tests. For simplicity of interpretation and because the Likert item data were not part of a validated Likert scale, the items were dichotomised to form two categorical variables “disagree” and “agree”, with responses 1 and 2 combined to form “disagree” and responses 4 and 5 combined to form “agree”. To determine if there was a significant difference in levels of agreement between Test 1 and Test 2 for these four questions a Pearson’s chi-squared (χ^2) test was performed.

3.3.2 Analysis two

For the adaptive test (Test 2) only, there were five statements specifically related to participant satisfaction with the test (adaptive statements A1-A5, Table 5). The analysis probed whether there were any significant differences in agreement for students studying in different countries and stages of study. In this analysis, the data was also dichotomised as in analysis one. In order to ascertain if country or stage of study or a combination of the two would predict the response frequency of “agree” multiple logistic regressions with “agree” as the categorical outcome variable were performed on the data. A Bonferroni correction was applied due to multiple comparisons being performed simultaneously (Armstrong, 2014).

To investigate whether experience with progress testing influenced students’ perception of the adaptive test, universities that regularly use progress testing and those that were new to progress testing were split into separate groups, and logistic regression with “agree” as the outcome variable was performed on this data.

All analysis was done using R open-source software package version 4.0.3 (R Core Team, 2020).

4 Findings

4.1 Descriptives

The overall response rate was 57% (Table 1, Tests 1 and 2), with a total of 458 participants responding to the survey in Tests 1 and 2, giving response rates of 30.0% and 98.0% respectively. This difference is based on the fact that remote delivery of Test 1 was very novel for students at the time, and the survey link was sent to students after the exam, while in Test 2 survey statements were incorporated into the QuizOne® and students responded immediately after the test.

In Tests 1 and 2, 73% of participants declared being fluent in English, which includes 34% of responders who declared English as their first language. Approximately 4.5% of participants declared being diagnosed with a learning disability (Table 2). Since the assessment statements and platform relied on English language, self-perception of the language proficiency was an important factor affecting subsequent answers to specific questions. Despite differences in stages, programmes, and to some degree language, there was a high level of satisfaction

Table 1 The number of participants who responded to the survey from each country and the response rate for Tests 1 and 2

Partner country	Finland	Portugal	Poland	Netherlands	UK	Total	Response rate %
Test 1 Number of responses to the survey using Qui-zOne®	22	40	21	39	18	140	30.0
Test 2 Number of responses to the survey using Qui-zOne®	11	49	34	85	139	318	98.0
Tests 1 + 2	33	89	55	124	157	458	57.0

Table 2 Participant demographic and language proficiency. Medical schools participating in the survey are categorized based on their familiarity with progress testing. Demographic characteristics of survey respondents are represented by the percentage of females, individuals in senior years (Stages 3–6), and those self-reporting a learning disability diagnosis. Self-reported language proficiency levels are also included. All data is presented as a percentage of survey responses

Demographic and language proficiency Test 1 and Test 2, % of survey responses	Universities with no progress testing			Universities using progress testing		Total %
	Finland	Portugal	Poland	Netherlands	UK	
Female	45	65	40	75	64	63
Diagnosed with learning disability	12	3	3.6	6	2.5	4.5
English as first language	6	3	25	9	72	34
Competent in English (self-reported *)	100	99	100	99	100	99.6
Fluent in English (self-reported **)	52	72	64	59	93	73
Senior years (Stages 3–6)	67	70	74	78	43	63

* "Yes" to the statement: I can understand texts that consist mainly of high frequency every day or job-related language. I can understand the description of events, feelings and wishes in personal letters

**"Yes" to the statement: I can read with ease virtually all forms of the written language, including abstract, structurally, or linguistically complex texts such as manuals, specialised articles, and literary works

with most aspects of the assessment. Table 3 shows the combined percentage of agree and strongly agree from answers in Tests 1 and 2 on 10 statements (S1-10). Eighty-two percent of all participants agreed that they obtained clear instructions for remote delivery of the tests (S1) and 80% agreed that they were clear about what was expected from them during the test (S2). Some significant variations between countries were observed with students from Poland and the Netherlands overall reporting the most satisfaction with the organisation and delivery of the online tests (S1-S5). Overall, 83% of all participants agreed that they were satisfied with the remote delivery of the tests (S3) and 85% agreed that they were satisfied with the practical organisation of the online tests (85%) though there were some local variations. A slightly lower proportion of participants (79%) agreed

Table 3 Combined participants' responses to ten statements that are common to both tests (Test 1 and Test 2). Summary response rate (% of agree and strongly agree) to 10 statements asked in both tests

Statements 1–10 in Test 1 and Test 2	Partner country % agree & strongly agree					Total
	Finland	Portugal	Poland	Netherlands	UK	
S1. Before taking the test, I received clear instructions on how to take the test remotely	76.0	78.0	89.0	91.0	77.0	82.0
S2. Before taking the test, it was clear to me what was expected of me	71.0	64.0	91.0	96.0	76.0	80.0
S3. I am satisfied with the remote delivery of the test	79.0	76.0	92.0	88.0	80.0	83.0
S4. I am satisfied with help I received during the remote delivery of the exam	79.0	68.0	87.0	93.0	74.0	79.0
S5. I am generally satisfied with the practical organisation of this test	70.0	79.0	92.0	94.0	82.0	85.0
S6. I think the software for this test was easy to use	88.0	79.0	89.0	99.0	79.0	86.0
S7. I think the questions were clearly formulated	70.0	80.0	85.0	67.0	57.0	68.0
S8. I think the quality of the items was correct	61.0	75.0	74.0	73.0	61.0	68.0
S9. I think the level of the test was appropriate for the stage of my study	39.0	45.0	53.0	50.0	49.0	48.0
S10. I think the length of the test was appropriate	58.0	54.0	66.0	87.0	72.0	70.0

that they were satisfied with the help they received during the tests (S4). Most of the student participants were satisfied with the practical organization of the test (overall 85%, S5), except in Finland where satisfaction was lower, which may be linked to the generally lower number of participants from this country. There was a high level of satisfaction related to the software being easy to use (overall 86%, S6), with the highest satisfaction from the Netherlands, Poland, and Finland, (Table 3). Opinions varied when participants referred to exam questions and assessment structure. Most participants agreed that questions were correct and clearly formulated (S7 and S8) with the lowest satisfaction among UK based students (all of whom had previous experience with progress testing assessments). In statement 9, whether the level of the test was appropriate for the students' stage of study and when considering answers from both tests combined the satisfaction was lower among all countries. The statement falls under the test content category and was analysed separately between Test 1 and 2 (see Fig. 1 and Table 4).

Overall satisfaction related to the length of the test (S10) was high (70%), although varied between countries from 54 to 87%, with the highest satisfaction

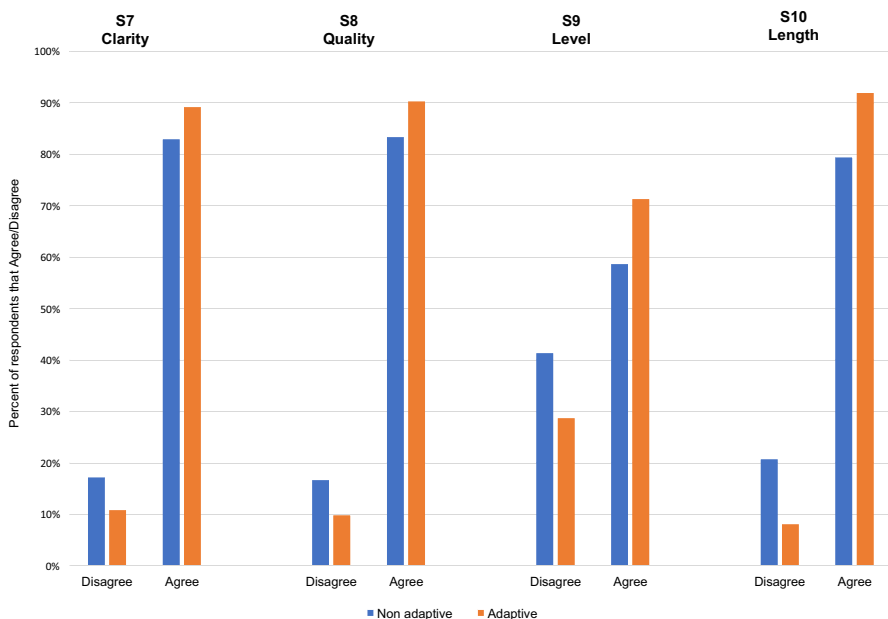


Fig. 1 Students' perception of the test clarity, quality of items, appropriate level of the test for stage, and the test length (S7-10) in Tests 1 and 2. Bar chart showing the percentage of respondents that agree or disagree with the four questions posed after the adaptive (orange, right bar in each set) and non-adaptive tests (blue, left bar in each set). In both tests, the combined agree and strongly agree answers to S7-10 are much higher than combined disagree and strongly disagree. In non-adaptive test, students were less satisfied with the level and the length of the test despite the overall high level of satisfaction

Table 4 A χ^2 analysis of the difference in responses to the four statements, S7-S10. The analysis includes the size of the effect between Test 1 (non-adaptive) and Test 2 (adaptive). Statistical differences were observed in statements 9 and 10 between tests

	χ^2	<i>df</i>	<i>N</i>	<i>p</i>	<i>phi Φ</i>	Effect size
S7. Clarity	2.68	1	351	0.102	0.087	ns
S8. Quality	3.38	1	353	0.066	0.098	ns
S9. Level	5.09	1	320	0.024*	0.126	small
S10. Length	12.11	1	376	0.001**	0.179	small

from countries that use progress testing routinely. Since statements 7–10 rely on the type of assessment, non-adaptive vs adaptive, we performed additional analyses on those.

4.2 Content of the tests

Figure 1 shows students' perception of the test clarity, quality of items, appropriate level of the test, and the test length (S7-10) in Tests 1 and 2. A high level of satisfaction is seen in both tests. χ^2 analyses identified significant differences in responses between Test 1 and Test 2 for statements 9 and 10, but not for statements 7 and 8 (Table 4). Whilst the results for S9 and S10 showed significant differences between the adaptive and non-adaptive tests the effect sizes were small (Cohen, 1988).

4.3 Analysis of statements specific to the Adaptive Progress Test (Adaptive, A1-5)

A logistic regression with “agree” as the outcome variable showed, that once a Bonferroni correction was applied, no statistically significant main effect for country or stage or interaction of country and stage across questions A1, A2 and A3. Some differences in stage were observed in statements A4 and A5 but these did not hold across all countries. In addition, no significant differences in students' perceptions were detected in all statements A1-A5 between students from universities routinely using progress tests and those institutions that do not, Table 5. However, most of the participants felt that the adaptive progress test is a good measurement of their knowledge and increases motivation for learning (A1-2). This opinion was shared irrespective of stage or previous experience with progress testing.

Moreover, Fig. 2 shows detailed Likert responses to statements A1-A5 suggesting that while the levels of agree and strongly agree (4 and 5 in Likert scale) may not be significant, there is a large proportion of participants who provided a neutral response. When asked whether the assessment was appropriate for the level of participants' knowledge to date (A3-5) students were less likely to be decisive in their answers (neutral A3 -36%; A4—34%, A5—31%, Fig. 2).

Table 5 The percentage of respondents that agreed with the statements A1 to A5 for Test 2 (adaptive test). Participants from Finland, Portugal and Poland had no prior experience with progress testing. Participants from UK and Netherlands were familiar with progress testing. The logistic regression analysis and significance levels show no statistical differences in responses irrespectively of prior familiarity with progress testing. 62% and 54% of participants agreed that the adaptive progress test is a good measure of their knowledge and increased motivation for study

Statements related to the adaptive test	Agree (%)				Logistic regression coefficient	p-value	Sig
	Universities with no progress testing	Universities with progress testing	Total				
A1. From this experience, I think the adaptive progress test is a good measurement instrument for my knowledge	64.5	61.7	62.5	0.2112	0.820	ns	
A2. I feel more motivated to study after the adaptive progress test compared with other knowledge tests I have experienced in my programme of study	59.8	51.8	54.2	1.3341	0.216	ns	
A3. In my opinion, the structure of the questions was well related to my level of knowledge	52.1	43.3	46.0	0.0397	0.960	ns	
A4. The questions I had to answer matched my level of knowledge	45.1	37.4	39.6	0.2313	0.780	ns	
A5. The questions I had to answer were representative of the level of knowledge I have gained in my studies up until now	48.3	41.1	43.2	0.2548	0.746	ns	

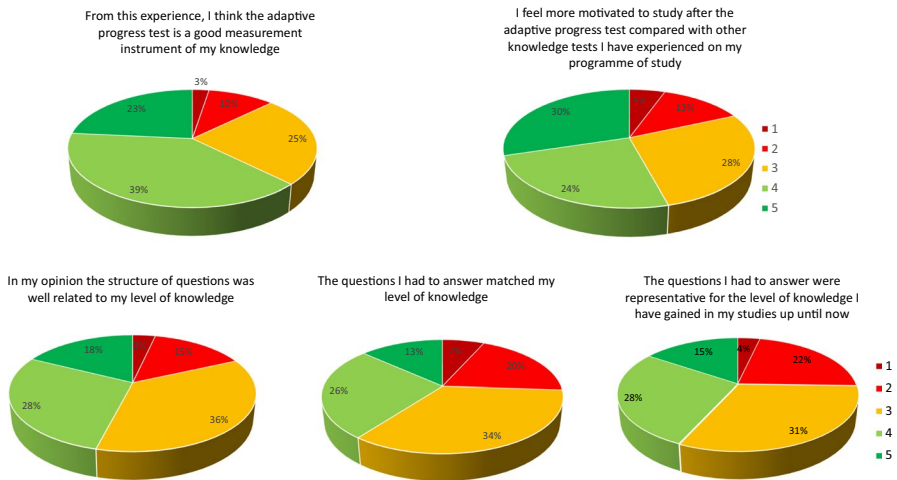


Fig. 2 All responses to A1-A5 statements provided after the adaptive progress test. Colour legend corresponding to Likert from 1–5 is on the right side of each row. A high level of “neutral” (yellow, the third triangle from the top in each pie chart) responses were observed in all statements. Overall, there is a much higher level of acceptability (green, medium size triangles on the left of each pie chart) of adaptive test across all questions when compared to students being less satisfied (red, smallest triangles on the top right of each pie chart)

5 Discussion

Previously we showed that adaptive progress testing is a reliable measurement of knowledge across different participants’ abilities irrespective of country and stage of training (Rice et al., 2022). Here we analysed factors that can potentially influence students’ acceptance of online adaptive progress testing. Our results suggest that there is a high level of student satisfaction with online remote delivery of progress tests, including the quality of instruction, acceptability of the QuizOne® platform, and the length and organisation of the test.

The acceptability of adaptive testing was equivalently high in students from different European countries with different curricula and languages. Overall satisfaction levels with the organisation and delivery of online tests were high regardless of previous experience with progress testing. Satisfaction with the help students received during the tests was somewhat lower with more variability between schools. Remote delivery and remote proctoring of tests were novel in all participant schools, for both students and staff, the latter acting as remote invigilators. Although the project consortium did provide training materials for invigilators and students, the variations in satisfaction in this area likely relate to the confidence and responsiveness of remote invigilators, as well as local Wi-Fi and technology. More research into the reasons why there are local variations in the acceptability amongst test takers of remote proctoring is required.

The idea that CAT links students’ abilities with true knowledge mastery provides an attractive assessment development that can lead to increased students’ motivation for learning and in turn improve their performance (Vollmeyer & Rheinberg,

2006). In this project, the CAT algorithm selected exam questions based on their difficulty aligning items with students' level of knowledge. In line with previous findings suggesting that CAT improves learner motivation and engagement when compared with paper-based tests (Martin & Lazendic, 2018), more students in our study agreed that the adaptive testing increased their motivation for learning, and that it was a good measure of their knowledge. However, this effect did not show statistical significance. When analysing detailed responses to questions A1-A5, it was evident that significant proportion of students provided a neutral response. As indicated by Nadler and colleagues (Nadler et al., 2015), this may represent a lot of meanings from being "unsure" or "undecisive" to a genuine mid-point response indicating either "both" or "neither". In the case of the adaptive test, where the algorithm selects questions based on previous answers, the neutral response should not be ignored as it can point towards an important self-perception of being unable to decide. A neutral response may indicate, for example, that the student did not have the insight to discern whether the test successfully aligned to their level of knowledge or not. There is a lot of debate around mid-scale response (DeMars & Erwin, 2005), and as such there is a case for arguing that neutral responses should be analysed within the particular context of the survey. However, the only way a neutral opinion can be classified as being more or less satisfied would require associated qualitative comments from students, which we did not have in this study.

Interestingly, students favoured the adaptive versus non-adaptive test in terms of the length of the test even though the adaptive test length for this study was fixed to 125 items (the same length as the non-adaptive test). Given that the adaptive test algorithm adjusts the level of difficulty to the individual student in real time the perception that adaptive testing promotes engagement with the test through question relevance appears to lead to positive perceptions of the length of the test. Nevertheless, students from schools who were experienced with progress testing (UK and Netherlands) were significantly more likely to agree that the length of the progress tests (125 items) was appropriate, whereas students from schools new to progress testing were less accepting of the test length. Furthermore, students from schools with progress testing experience were significantly less likely to agree that the exam questions were clearly formulated (S7) than their peers from schools that do not routinely use progress testing. Previous experience with progress testing and English language proficiency likely played a role in the acceptability of the question formulation within the assessment.

Using a remote proctored platform for delivery was also successful and enabled the tests to be administered during COVID-19 lockdowns. In line with recent studies (Jaap et al., 2021) on remote delivery of exams in medical education, we have experienced very few technical issues, with the most frequent issues being wifi connectivity and camera access.

Using adaptive testing sequentially as a progress test allows personalised assessment since the difficulty of the test adjusts to the ability of the learner (Wainer et al., 2000; Rice et al., 2022). In this way, progress testing can have utility not just as a summative assessment, but also as an assessment for learning, since students are more likely to engage when being tested on areas of medical knowledge which are appropriate for their stage of learning with a test that adjusts difficulty in an adaptive

algorithm. The self-perception of students on given assessment needs to be linked to understanding how feedback can help in knowledge growth. It should be noted that in our project most participants outside of the UK schools were in stages 3–6 of their medical programmes. If there were more students from earlier stages with less applied medical knowledge at the level of the progress test, we would expect the impact of adaptive testing on the perceived appropriateness of the level of the test to increase. Whilst the adaptive testing should in principle benefit students in early stages, we did not observe a significant difference in students' perceptions between early (years 1 and 2) and later stages (years 3–6). The psychometric properties of progress tests have been found to be less robust at early stages of the medical curricula compared to later stages (Ricketts et al., 2009), but adaptive testing has been shown to have good reliability when measuring the performance of students in the early stages of medical school (Rice et al., 2022).

5.1 Implications

To our knowledge, published literature comparing student perceptions of non-adaptive vs adaptive assessment is very limited. Therefore, this study is important for future development of knowledge assessment in various programmes. The scale of our study involving students from across Europe, medical programmes and stages and collecting their opinions provides a significant insight into wider acceptability of adaptive progress testing in general. Our study adds to the existing knowledge base around adaptive testing applications in terms of increased motivation for learning and engagement with the test material. Future research into longitudinal outcomes will be interesting to build on these findings. Whilst a frequently cited advantage of CAT is the ability to reduce test length (as the adaptive algorithm enables convergence to students' true knowledge level relatively quickly), this is not necessarily a desirable solution for progress testing from a wide curriculum blueprint in terms of the coverage of the test and the ability to enable the provision of detailed feedback to students on a test-by-test basis. However, the use of asynchronous adaptive progress testing (progress tests anytime, anywhere), perhaps even on-demand, could indeed lead to knowledge tests becoming more frequent and shorter, maintaining adequate blueprint coverage and optimizing test reliability by lowering measurement error for students even in early academic years. More research into the acceptability, among both students and medical faculty, of adaptive progress testing used in this way is required.

5.2 Limitations

Moving from the face-to-face test delivery platform (TestLife®) used in the pilot to the remote delivery platform (QuizOne®) for Test 1 and Test 2 potentially introduced a confounding factor that affected the student response, therefore our analysis included only data from the two remotely delivered tests.

In addition, it appears that some of the participants did not notice that a survey followed Test 1, the number of participants was high, but only 30% of those students responded to all questions/statements in the survey. Learning from this experience the survey for Test 2 was seamlessly attached to the test, which resulted in a 98% response rate.

6 Conclusions

Medical programmes cover a large number of learning outcomes that require students to utilise effective ways of self-directed learning. Progress testing promotes deep learning assessing students toward the final programme outcomes, but it relies on specific standards for each stage. Adapting the assessment to students' knowledge may provide a unique opportunity for detailed feedback increasing students' motivation for learning. Adaptive progress testing is well-accepted by students across multiple countries, universities, and stages of study, regardless of whether they are familiar with progress testing or not and it is potentially preferred by students to conventional non-adaptive testing.

Acknowledgements The authors wish to thank Annemarie Camp and Evelien Neis, who project managed the consortium development and test administration, and all members of the OAIPT consortium who have contributed to the outputs of the project to date: Jagiellonian University, Poland (Mateusz Rubinkiewicz); Maastricht University, The Netherlands (Carlos Colares, Joyce Moonen-van Loon, Cees van der Vleuten); Medical University of Lodz, Poland (Janusz Janczukowicz, Katarzyna Janusz, Paulina Sobieranska); University of Exeter, UK (Kevin Brandom, Adrian Freeman, Jonathan Wyatt); University of Helsinki, Finland (Otto Helve, Mika Laitenan, Johanna Louhimo); University Medical Center Groningen, The Netherlands (Debbie Jaarsma, Bram Jacobs, Michiel Katoele, Ally van Hell); iCognitus4ALL, Portugal (Nuno Santos).

Funding This study was supported by the Erasmus+ grant 2018-1-NLOI-KA203-038925.

Data availability The anonymized datasets used for all analyses in this study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval This research has been approved by the Netherlands Association for Medical Education (NVMO) Ethical Review Board, Maastricht, NERB dossier number: 2019.4.2 (25.06.2019).

Financial interests The authors declare they have no financial interests to disclose.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.








References

- Ali, K., Coombes, L., Kay, E., Tredwin, C., Jones, G., Ricketts, C., & Bennett, J. (2016). Progress testing in undergraduate dental education: The Peninsula experience and future opportunities. *European Journal of Dental Education*, 20(3), 129–134. <https://doi.org/10.1111/eje.12149>
- Ali, K., Zahra, D., Tredwin, C., Mcilwaine, C., & Jones, G. (2018). Use of progress testing in a UK dental therapy and hygiene educational program. *Journal of Dental Education*, 82(2), 130–136. <https://doi.org/10.21815/JDE.018.015>
- Armstrong, R. A. (2014). When to use the Bonferroni correction. *Ophthalmic and Physiological Optics*, 34(5), 502–508. <https://doi.org/10.1111/opo.12131>
- Boud, D., & Soler, R. (2015). Sustainable assessment revisited. *Assessment and Evaluation in Higher Education*, 41, 1–14. <https://doi.org/10.1080/02602938.2015.1018133>
- Brown, G. T. L., & Hirschfeld, G. H. F. (2008). Students' conceptions of assessment: Links to outcomes. *Assessment in Education: Principles, Policy & Practice*, 15(1), 3. <https://doi.org/10.1080/09695940701876003>
- Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Lawrence Erlbaum Associates, Publishers.
- Collares, C. F., & Cecilio-Fernandes, D. (2019). When I say ... computerised adaptive testing. *Medical Education*, 53(2), 115–116. <https://doi.org/10.1111/medu.13648>
- Common European Framework of Reference for Languages (CEFR) (2018). *The CEFR Levels*. [online] Common European Framework of Reference for Languages (CEFR). Retrieved June 2018 from <https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions>
- DeMars, C. E., & Erwin, T. D. (2005). Neutral or unsure: Is there a difference? Poster presented at the annual meeting of the *American Psychological Association*, Washington, DC, <https://commons.lib.jmu.edu/gradpsych/30/> Accessed Jan 2023.
- Devine, O. P., Harborne, A. C., & McManus, I. C. (2015). Assessment at UK medical schools varies substantially in volume, type and intensity and correlates with postgraduate attainment. *BMC Medical Education*, 15, 146. <https://doi.org/10.1186/s12909-015-0428-9>
- Griff, E. R., & Matter, S. F. (2013). Evaluation of an adaptive online learning system. *British Journal of Educational Technology*, 44(1), 170–176. <https://doi.org/10.1111/j.1467-8535.2012.01300.x>
- Ho, G. W. K. (2017). Examining perceptions and attitudes: A review of likert-type scales versus Q-Methodology. *Western Journal of Nursing Research*, 39(5), 674–689.
- Hoffmeyer-Zlotnik, J. H. P. (2016). Standardisation and Harmonisation of Socio-Demographic Variables (Version 2.0). GESIS Survey Guidelines. Mannheim, Germany: GESIS – Leibniz Institute for the Social Sciences. https://doi.org/10.15465/gesis-sg_en_012
- Jaap, A., Dewar, A., Duncan, C., et al. (2021). Effect of remote online exam delivery on student experience and performance in applied knowledge tests. *BMC Medical Education*, 21, 86. <https://doi.org/10.1186/s12909-021-02521-1>
- Martin, A. J., & Lazendic, G. (2018). Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. *Journal of Educational Psychology*, 10(1), 27–45. <https://doi.org/10.1037/edu0000205>
- Mohebi, L., & Bailey, F. (2020). Exploring Bem's Self perception theory in educational context. *Encyclopaedia*, 24(58), 1–10. <https://doi.org/10.6092/issn.1825-8670/9891>
- Nadler, J. T., Weston, R., & Voyles, E. C. (2015). Stuck in the middle: The use and interpretation of mid-points in items on questionnaires. *The Journal of General Psychology*, 142(2), 71–89. <https://doi.org/10.1080/00221309.2014.994590>
- R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved June 2021 from <http://www.r-project.org/index.html>
- Rice, N., Pêgo, J. M., Collares, C. F., Kisielewska, J., & Gale, T. (2022). The development and implementation of a computer adaptive progress test across European countries. *Computers & Education: Artificial Intelligence*, 3, 10083. <https://doi.org/10.1016/j.caeai.2022.100083>
- Ricketts, C., Freeman, A., & Coombes, L. (2009). Standard setting for progress tests: Combining external and internal standards. *Medical Education*, 43, 589–593. <https://doi.org/10.1111/j.1365-2923.2009.03372.x>
- Ross, B., Chase, A. M., Robbie, D., Oates, G., & Absalom, Y. (2018). Adaptive quizzes to increase motivation, engagement and learning outcomes in a first year accounting unit. *International Journal of Educational Technology*, 15, 30.

- Sullivan, G. M., & Artino, A. R., Jr. (2013). Analyzing and interpreting data from likert-type scales. *Journal of Graduate Medical Education*, 5(4), 541–542. <https://doi.org/10.4300/JGME-5-4-18>
- Van der Vleuten, C., Verwijnen, G., & Wijnen, W. (1996). Fifteen years of experience with progress testing in a problem-based learning curriculum. *Medical Teacher*, 18, 103–109. <https://doi.org/10.3109/01421599609034142>
- Vollmeyer, R., & Rheinberg, F. (2006). Motivational effects on self-regulated learning with different tasks. *Educational Psychology Review*, 18, 239–253. <https://doi.org/10.1007/s10648-006-9017-0>
- Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (2000). *Computerized adaptive testing: A primer*. Mahwah, NJ: Lawrence Erlbaum Associates (2nd ed).
- Wrigley, W., Van der Vleuten, C. P., Freeman, A., & Muijtjens, A. (2012). A systemic framework for the progress test: Strengths, constraints and issues. AMEE Guide No. 71. *Medical Teacher*, 34, 683–697. <https://doi.org/10.3109/0142159X.2012.704437>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Jolanta Kisielewska^{1,2}  · Paul Millin^{1,2}  · Neil Rice^{2,3}  ·
Jose Miguel Pego^{2,4,5}  · Steven Burr^{1,2}  · Michal Nowakowski^{2,6}  ·
Thomas Gale^{1,2} 

✉ Jolanta Kisielewska
jolanta.kisielewska@plymouth.ac.uk

- ¹ Peninsula Medical School, Faculty of Health, University of Plymouth, Plymouth, UK
- ² European Board of Medical Assessors, ERASMUS+OAIPT Project Team, Maastricht, Netherlands
- ³ College of Medicine and Health, University of Exeter Medical School, University of Exeter, Exeter, UK
- ⁴ Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho, Braga, Portugal
- ⁵ ICVS/3B's-PT Government Associate Laboratory, Braga, Guimarães, Portugal
- ⁶ Faculty of Medicine, Jagiellonian University Medical College, Krakow, Poland